nature

26 April 2001 Volume 410 Issue no 6832

Next steps against AIDS

The South African government at last has an opportunity to fight AIDS with drugs at reduced prices without fear of legal action. The country has the capacity to make significant progress. The government should make use of it.

The withdrawal last week of the case against the South African government by the country's Pharmaceutical Manufacturers Association and 39 drug companies (see page 1013) should be widely welcomed. After more than three years of attrition, it is high time that the industry and government begin to work together to address mother-to-child transmission of HIV, as well as the treatment of South Africa's estimated half a million AIDS sufferers.

The ball now lies in the court of the South African government, which can no longer remain paralysed over AIDS policy. The reaction of its health minister, Manto Tshabalala-Msimang, has been disappointing. Her statement that the government has no intention of importing antiretrovirals in the near future raises suspicions that the government refuses to confront the realities of the country's predicament. The same is true of her claims that treating opportunistic infections and providing adequate nutrition will permit AIDS sufferers to "function adequately".

No one is under any illusions that the next steps will be easy: South Africa is estimated to have the greatest number of HIV-positive people — 4.7 million — of any nation. But its medical infrastructure is, by the standards of the developing world, relatively good, and most of its population is urban and has access to treatment. In many parts of the country the capacity to administer the drugs adequately and effectively is in place, despite government claims to the contrary. It would be tragic if AIDS sufferers and unborn children were not to benefit from the offers of cheap or free drugs.

In most cases, the offers from manufacturers are confined to the government sector. Drug companies would do well to consider the example of GlaxoSmithKline in extending its offer of reduced prices to non-governmental organizations and private employers who offer care to their staff through clinics at their own workplace. But this can nonetheless serve only a fraction of the people who could potentially benefit from concerted government action.

Tshabalala-Msimang is in Europe next month to negotiate with the drug companies on the regulations regarding the South African legislation. Reportedly, the South African government has undertaken to confine parallel importation of drugs to branded versions under patent in South Africa rather than import generic copies, and to issue compulsory licences in compliance with the Patents Act. This should reduce disparities in drug prices between markets, while enshrining intellectual property rights. Such a compromise should hopefully be reflected in good legislation, but that will serve little purpose unless it is applied.

The end of the South African suit will switch attention to the possible action being brought against Brazil within the World Trade Organization (WTO), backed by the Pharmaceutical Research and Manufacturers of America. The association says Brazil is flouting the treaty relating to intellectual property rights by making cheap copies of patented AIDS drugs. The case will soon go before the WTO, and hopefully, as in South Africa, the outcome will be a realistic compromise. Brazil is an example of how a relatively poor country can treat AIDS — admittedly significantly less prevalent than in South Africa — if it has access to cheaper generic drugs: AIDS deaths have halved since the government began providing cut-price treatment. The South African government should seize the opportunity to follow that example, with its own hard-won compromise from industry behind it. ■

Astronomy, goddesses and knowledge

Conflict between science and spiritual traditions in Hawaii can be overcome by common interests.

Astronomers might once have considered themselves lucky to work so close to heaven. Now it seems a liability. High mountaintops in Arizona and Hawaii have become battlegrounds between scientists straining to improve their view of the Universe and traditional peoples who have scaled the same heights for centuries, seeking other kinds of knowledge.

A neutral observer might see room at the top for both. But the increasingly common clashes over sacred mountains such as Mauna Kea (see page 1015) are made vastly more difficult by the mutual suspicion and hurt feelings that are endemic to the politics of culture.

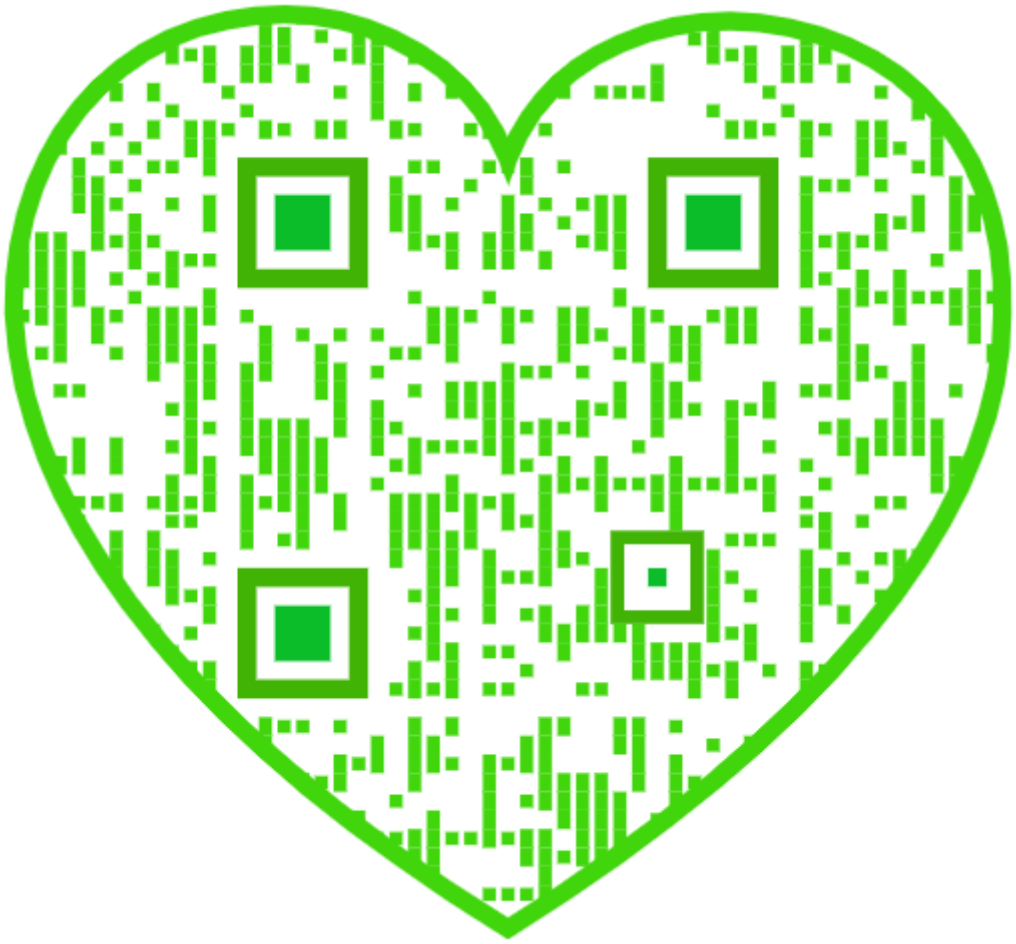
Hawaii, the last state to join the United States in 1959, has seen a resurgence of pride in traditional Polynesian culture in recent years. In 1993, the US government took the extraordinary measure of issuing an official apology to native Hawaiians for the overthrow of the Kingdom of Hawaii 100 years earlier. Just this month, the state's two US senators, mindful of sentiment back home, reintroduced a bill in Congress to establish a quasi-sovereign Native Hawaiian government.

Having enjoyed mostly good relations with the 'Aloha State' for more than 30 years, scientists, who come almost exclusively from the mainland, are being swept along by this rising cultural tide. If Hawaiian traditionalists feel bullied by astronomers and their big, expensive machines, the scientists have had their own feelings hurt, and resent being lumped in with shopping-mall builders and golf-course developers as symbols of Western crassness. Many astronomers believe that they, too, are on a quest for truth and beauty.

So what is to be done? NASA, the US space agency (which, ironically, rarely dabbles in ground-based astronomy, and must wish it never had in this case), is following the right course. Frustrating as it surely is to be castigated in town meeting after town meeting, NASA managers should remain respectful of Hawaiian traditions, and continue their 'constructive engagement'. Astronomers familiar with the Mauna Kea issue admit that scientists could be more sensitive to the fact that they're working in someone else's church.

The native Hawaiian opposition, for its part, should recognize good-faith gestures when they're offered. And they should ask themselves: do we really want to drive these telescopes — which reveal a grandness to nature that our ancestors would have appreciated — off the mountain for ever? ■

© 2001 Macmillan Magazines Ltd



Foot-and-mouth
Tests for vaccinated animals offer means to track disease
p1012

Earth Day
US administration bolsters its green credentials
p1014

Culture clash
Keck telescope upgrade delayed on 'sacred' mountain
p1015

Blast off
India celebrates successful rocket launch at last
p1017

Russia needs help to fend off potato famine, researchers warn

Quirin Schiermeier, Munich

Aggressive strains of potato blight could trigger a potentially catastrophic potato famine in Russia, scientists have warned.

The researchers want an international effort to provide Russian potato-growers with fungicides and seed varieties that are resistant to virulent late blight pathogens, which last year destroyed more than 15% of the country's total crop.

"The situation is worrying, and it could quickly worsen," warns Kandukuri Raman, a plant scientist at Cornell University in Ithaca, New York, and head of the Cornell–Eastern Europe–Mexico (CEEM) project on control of potato late blight.

Potato late blight, which is caused by the fungus *Phytopthora infestans* and originated in Mexico, is one of the world's most devastating crop diseases, with annual losses and fungicide costs estimated at US$3 billion.

Potatoes are a vital subsistence crop grown by poor Russians who cannot afford to buy vegetables or meat. Russia is the world's second-largest potato producer, after China.

Raman fears that the state of its economy means that Russia will not be able to control the disease without outside support. He points out that Russia has no strong potato-breeding programme aimed at developing varieties resistant to late blight.

Spores of *P. infestans* constantly reproduce and take on characteristics that enable them to survive and spread. The new, aggressive strains proliferate freely and can survive in the soil over winter, attacking potatoes in summer and destroying the harvest.

Late blight caused the Irish potato famine in the 1840s, leading to the death of a million people and a huge wave of migration. But fungicides have been available since the late nineteenth century, and have fought off even the more virulent strains across Europe. However, late blight remains a major threat to small farmers and householders in Russia, who lack access to suitable fungicides.

"Pesticide-resistant and virulent strains of *P. infestans* have appeared in the major potato-growing regions in Russia," says Patrick Russo, a plant geneticist at the University of Helsinki in Finland and a technical adviser on control of late blight in Russia. Even types of potato that showed some resistance to late blight "are no longer resistant to the newer strains", he adds.

CORBIS

At risk: Russia's potatoes are vulnerable because their growers lack access to fungicides.

The CEEM project aims to help distribute seeds of resistant potato varieties to Russian smallholders, but has limited resources. It is also collaborating with Polish and Russian scientists to test wild potato species, and to study the molecular biology of the interaction between *P. infestans* and the potato.

Much of the work is being done at the N. I. Vavilov All-Russian Scientific Research Institute of Plant Industry in St Petersburg, which holds one the world's oldest and largest potato collections.

The US Department of Agriculture and the CEEM project are sponsoring a June workshop in Warsaw on potato late blight. Experts from ten countries will discuss the use of plant genetics and traditional breeding methods to develop broad-based resistance, and explore the possibility of a programme to forecast the arrival of different strains of the pathogen. ■

http://www.cals.cornell.edu/dept/plantbreed/CEEM

Smithsonian closure plan under fire

Josette Chen

Leading Republicans have attacked plans by the Smithsonian Institution to close its Conservation Research Center (CRC) at Front Royal, Virginia (see *Nature* 410, 727; 2001).

Congressman Frank Wolf (Republican, Virginia), a member of the House Appropriations Committee which controls the government's share of the Smithsonian budget, said: "I am extremely disappointed and angry with the Smithsonian. This decision — and the way it has been announced — is shortsighted and ill-advised."

Senator John Warner (Republican, Virginia), who visited the facility on 21 April, said the closure "does not reflect a consensus of views in the research community".

In a letter to the Smithsonian's secretary Larry Small, Sherwood Boehlert (Republican, New York), chairman of the House Science Committee, said: "To close the CRC or curb its capabilities would be a blow to the Smithsonian and its reputation, as well as to the nation and the world."

The closure will be considered on 7 May by the Smithsonian's governing board. It must then be approved by Congress, where influential members favour the centre's continued operation.

At a meeting with 500 museum staff on 17 April, Small discussed the Smithsonian's need to refurbish buildings as part of a proposed reorganization. But he did not detail any cuts in research programmes.

The Smithsonian has rejected plans from the Fish and Wildlife Service for joint management of the centre. But last week National Zoo director Lucy Spelman seemed to retract her earlier firm announcement of its closure. Chris Wemmer, the centre's associate director, says she referred to it as only a proposal. ■

© 2001 Macmillan Magazines Ltd

Standards needed for foot-and-mouth tests

Sally Goodman, Paris

Simple tests for foot-and-mouth disease that differentiate between vaccinated, infected and 'carrier' livestock could become available quickly if international agreement was reached on sensitivity and specificity, a meeting in Paris was told last week.

The International Scientific Conference on Foot-and-Mouth Disease brought together representatives of the UN Food and Agriculture Organization (FAO) and the main intergovernmental animal-health organization, the Office International des Epizooties (OIE), which covers 157 countries.

Delegates heard that such tests would be highly valuable to countries that want to maintain disease-free status and export livestock or meat after a vaccination programme or epidemic. The tests would make it far easier for countries to vaccinate animals, rather than slaughter them, as Britain has done, in response to outbreaks of the disease.

The diagnostic tests in question are based on enzyme-linked immunosorbent assay techniques. These can differentiate between antibodies in the blood against structural proteins found in whole viruses and against non-structural proteins, which are produced as the virus replicates. Because the vaccines tend to be produced from inactivated viral particles, vaccinated animals do not usually generate antibodies against the non-structural proteins and so can be identified by the assays.

CORBIS

Under examination: vaccination against foot-and-mouth disease could become more common.

"The tests are proving useful but they aren't yet 100% perfect," said Tony Garland, of Britain's Institute for Animal Health (IAH) in Pirbright, Surrey. "They are currently only suitable for diagnosis at herd level, not at the individual level." The IAH and the Istituto Zooprofilattico Sperimentale in Brescia, Italy, in collaboration with the Dutch vaccine company Intervet, hope to have their test on the market in September. Another commercial test has been produced by United Biomedical, a pharmaceutical company based in New York state.

But before they can be produced in sufficient quantities for mass screening or be accepted by the World Trade Organization as proof of disease-free status for export purposes, the tests need to be harmonized against a set of standard varieties of the disease virus. The complexity of the foot-and-mouth virus and its wide range of hosts makes this standardization hard to achieve.

The process could take between one and three years, according to Bernard Vallat, director-general of the OIE. Another problem is that the OIE's general assembly, which has the final vote on such standardization, meets only once a year.

At last week's conference, experts called for international concerted action to speed up the process. They proposed a list of recommended actions on foot-and-mouth disease to be put before the OIE's annual meeting in May.

"Steps should be taken on preventative measures in areas where the virus is endemic," said John Crowther, a member of a Vienna-based collaboration between the International Atomic Energy Agency and the FAO animal-health division.

The meeting also recommended special provision for emergency vaccination of endangered species and those carrying rare genetic material, without prejudice to the disease status of the host country, provided that the animals are physically separated from other populations. But little research has been done on the effectiveness of vaccines in rare and exotic species. ■

http://www.oie.int

Reef gets off the starting blocks

David Adam, London

Marine scientists in Scotland are poised to dump one and a quarter million concrete blocks onto the seabed to study the effects that artificial reefs have on sea-life.

Weighing over 40,000 tonnes and taking more than two years to build, the reef will help researchers to study how marine animals select and colonize habitats, as well as offering clues to how such giant constructions affect local sea currents and seabed sediment quality.

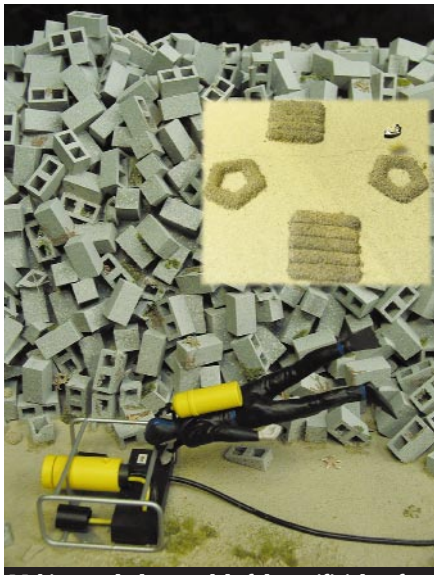
The project, which will cost £1.5 million (US$2 million), has now been granted a licence by the Scottish parliament. The first blocks will be dropped next month.

The reef is being built off the island of Lismore in Loch Linnhe, near Oban on the west coast of Scotland. Martin Sayer, a marine scientist with the Scottish Association for Marine Science (SAMS) laboratory at Dunstaffnage in Argyll, who is directing the scheme, says that most scientific research on artificial reefs has taken place on industrial installations such as breakwaters. "This will be different because its primary objective is experimental," he explains.

The reef will provide a platform for various SAMS experiments, he says, but will also be available to other research groups. A marine mammal team at the University of St Andrews has already arranged to use it to study porpoises.

The concrete blocks will be dumped in 24 separate reefs over an area the size of six football pitches. To test the effects of different reef and block shapes, half of the blocks will be solid, half will have holes, and both kinds will be used in square- and pentagon-shaped reef designs. ■

http://www.sams.ac.uk/sams/dml/projects/reef



Making a splash: a model of the artificial reefs to be created in Scottish waters.

© 2001 Macmillan Magazines Ltd

Japan speeds up mission to unravel genetic diseases

David Cyranoski, Tokyo

A sealed card is at the heart of Japan's bid to unmask the genetics behind various diseases including arthritis and asthma. With the opening of its new facility in Tokyo, RIKEN, the Institute of Physical and Chemical Research, plans to push Japan to the forefront of the global effort to interpret the human genome.

The facility will use its own sealed-card technology to map single-nucleotide polymorphisms (SNPs), the single-base variations that are scattered throughout the genome. SNPs are thought to be behind most genetic variations — from eye colour to disease susceptibility.

RIKEN's technology allows researchers to read up to 384,000 DNA samples per day, using much smaller blood samples than were previously required. The institute hopes the technique will allow it to match and even surpass existing efforts, such as the SNP Consortium of international drug companies and academic laboratories.

"The capacity they are trying to achieve will put them at the top of the heap," says Pui-Yan Kwok, a genetics researcher at Washington University School of Medicine in St Louis, Missouri.

The new facility will be managed by Yusuke Nakamura, who is also head of Tokyo University's Human Genome Center. This centre has already identified about 90,000 new SNPs within the Japanese population, and aims to reach 150,000 by next March.

Many diseases, such as cancer and diabetes, are associated with several genes. SNPs can be used to identify those genes — a SNP located close to a gene associated with a certain disease tends to be inherited with it, and so can act as a marker for the disease.

But some SNPs are found within the disease genes themselves, and it is these that Nakamura is most interested in. This is because they can help identify not just groups of genes, but also particular combinations of alleles — gene sequences with small variations — that are associated with the disease.

Nakamura plans to genotype 120,000 SNPs for each of 768 Japanese patients. This will mean analysing about 100 million SNPs, a mammoth task that he hopes to finish within a year.

"It is necessary to look simultaneously at hundreds of thousands of SNPs in an individual's genome and then compare these with other individuals — thousands or tens of thousands of them," says Nakamura.

RIKEN will achieve this kind of capacity by combining its sealed, 384-well card with a technique for amplifying DNA from blood samples, and an assay that can quickly genotype SNPs. Together these reduce the size of the sample required, reducing costs to "less than one tenth of the previous method", says Nakamura.

DAVID CYRANOSKI

Rapid response: Yusuke Nakamura says the sealed-card technology will speed up analysis.

The sealed card helps to speed up the entire process, explains Nakamura. Conventional cards have to be processed one by one — placed in a scanning machine that incubates the materials in the wells and catalyses the necessary reaction for detection. But RIKEN's cards can be dropped into hot water in large batches, allowing 1,000 cards to be dealt with each day.

As RIKEN ramps up its identification of SNPS, others are not sitting idle. The Whitehead Institute for Biomedical Research at the Massachusetts Institute of Technology is mounting an effort of similar size and scope to that at RIKEN. This will focus on American and European populations, says Eric Lander, director of the institute's Center for Genome Research. "I reckon that these will be the largest efforts in the world — and quite complementary," he says.

Nakamura says he has already found SNPs related to arthritis and asthma, and expects to find many more as his genotyping project develops. Like the SNP Consortium, RIKEN plans to make all of its findings freely available on the Internet. It will post the first 10,000 of the genotyped SNPs by May or June of this year, and put all of the genotyped SNPs information in a public database next year. ■

http://snp.cshl.org

http://snp.ims.u-tokyo.ac.jp

South Africa may keep door closed to generic AIDS drugs

Michael Cherry, Cape Town

Drug companies expect the South African government to hold off from importing cheap generic copies of their AIDS drugs following the withdrawal of their lawsuit against the government last week.

Although the 39 drug-makers withdrew their lawsuit unconditionally, officials of South Africa's Pharmaceutical Manufacturers Association say the government has privately undertaken to import only branded versions of drugs already under patent in South Africa.

The government's only public promise is to consult the drug companies when drawing up regulations relating to the section of the 1997 Medicines and Related Substances Control Amendment Act to which the companies objected in their lawsuit.

The government had previously said it would use either parallel imports or compulsory licensing — allowing domestic production of a patented drug — to obtain drugs at lower prices than offered by patent holders. It has not committed itself to withdrawing that intention.

Earlier last week, South Africa's Medicines Control Council announced that it had registered the anti-AIDS drug nevirapine for use in preventing the mother-to-child transmission of HIV. The drug's manufacturer, Boehringer Ingelheim, has agreed that any possible resistance problems can be monitored. The decision opens the way — in theory — for the drug to become available on prescription.

But AIDS activists' euphoria over the lawsuit might be short-lived. At a press conference following the decision, health minister Manto Tshabalala-Msimang said hopes should not be raised that anti-AIDS drugs would become available. She said their widespread use would be hampered by a lack of adequate infrastructure, high costs and drug resistance. ■

AP

Good news: AIDS activists celebrate in court as the drug companies drop their lawsuit.

© 2001 Macmillan Magazines Ltd

US administration tries to repair green image

Mark Schrope

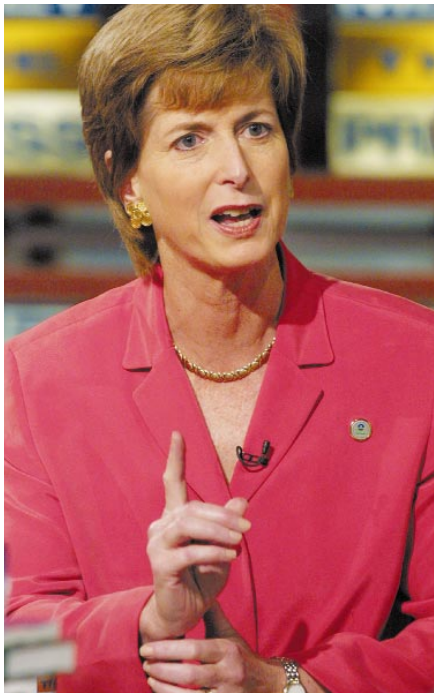
Following public criticism of its decisions on arsenic pollution and global warming, President George W. Bush's administration last week sought to bolster its green credentials with a barrage of pro-environment announcements.

On 18 April, the Environmental Protection Agency (EPA) said it was asking the National Academy of Sciences for a lightning study of the health effects of arsenic in drinking water. The agency wants to establish an appropriate standard for arsenic levels, having rejected one put forward by the Clinton administration (see *Nature* **410,** 503; 2001).

The EPA also said it would accept a number of other controversial environmental rules put in place during the last days of the Clinton administration. One would protect wetlands by closing a legal loophole that, under certain conditions, allows land developers to dredge and bulldoze these areas. Another will require more businesses to report how much lead they release into the environment.

The White House issued statements supporting the EPA's action, timed for the run-up to Earth Day, a public celebration of environmental progress in the United States, held on 22 April. But some business groups opposed the moves, and said that they plan to challenge the new rules in the courts.

On 19 April, the president announced that he would sign an international treaty, agreed under the Clinton administration, to phase out production and use of persistent organic pollutants such as polychlorinated biphenyls and the pesticide DDT.



Finger on the pulse: Christie Whitman promises new standards for arsenic in drinking water.

ALEX WONG / NEWSMAKERS

But environmentalists are sceptical. "This is basically a public relations ploy," says Elliott Negin, a spokesman for the Natural Resources Defense Council. He adds that he doubts whether the EPA will enforce the lead and wetlands rules with sufficient vigour.

White House spokesman Ari Fleischer denies that Bush is concerned about his environmental image, saying that the plans "represent a balanced approach to protecting America's environment".

To develop its arsenic standard, the EPA wants the National Academy to update its 1999 report on the health effects of arsenic in drinking water. Gary Carlson, a toxicologist at Purdue University in West Lafayette, Indiana, who was on the panel that produced the 1999 report, says that there is some value in updating it.

Carlson says the main issue is whether or not there is a threshold below which arsenic's effects are negligible. But scientists disagree on this fundamental point, which is why the 1999 study did not recommend a 'safe' level of arsenic, and why some argue that the decision to set a regulatory level is essentially political, rather than scientific.

EPA administrator Christie Whitman says that a new standard will be formulated by February 2002 and that compliance with it will be required by 2006, just as it would have been with the rule proposed by the Clinton administration. The EPA will also ask the National Drinking Water Advisory Council to examine economic issues related to arsenic reductions. ■

Tenure-track plan aims to end university inbreeding

Xavier Bosch, Barcelona

Spain has introduced a new career track for young researchers. The initiative aims to increase competition and eliminate the academic patronage that critics say hampers the development of Spanish science.

The Ramón y Cajal programme, named after the Nobel-prizewinning neuroscientist, will offer five-year contracts to 2,000 postdoctoral fellows over the next two years.

Conditions under the new programme — which will name its first 800 fellows this year — will be far better than under the existing Spanish postdoc recruitment system. In addition to a higher salary — Ptas 4.75 million (US$26,000), compared with Ptas 2.9 million — and a contract of five years instead of three, Cajal postdocs will be allowed to lead research projects of their own.

University departments and research centres wishing to join the programme must submit a seven-year plan, covering research and personnel, to prove they will be able to offer permanent positions afterwards. External panels of scientists will evaluate all postdocs' progress after two and four years.

In a clear effort to increase movement between centres, applicants must have spent 18 months working somewhere other than their place of application. The programme will lead to "an improved quality of research, easier mobility and less 'inbreeding'", said the Ministry of Science and Technology, announcing the programme last week.

The initiative marks "an important change in mentality", says Ramon Marimon, the science ministry's secretary of state. "It's a turning point for research careers in Spain as it introduces the concept of tenure track and, by opening up research centres' planning procedures to public view, it should guarantee the sustainable growth of research positions."

Eugenio Santos, director of the University of Salamanca's Cancer Research Institute, says that "for the first time" Spain will have a tenure-track system that is up to international standards of competition and evaluation. He says the programme should attract "a large pool of well-trained young Spanish scientists" who have completed postdoctoral periods abroad and now want to develop their research careers.

Jordi Petriz, a postdoc working on stem-cell research at the University of Barcelona, says he welcomes the "leadership and continuity" the system will offer. ■

© 2001 Macmillan Magazines Ltd

Astronomers bargain for use of 'sacred' site

Tony Reichhardt, Washington

Construction of a NASA-funded telescope array on Mauna Kea in Hawaii has been held up by wrangling between astronomers and Hawaiians over the use of the sacred mountain, already home to some of the world's most powerful telescopes.

The delay to the $50 million Keck Interferometer — now stretching into its seventh month — is being watched closely by astronomers, who fear that one of their premier viewing sites is becoming increasingly mired in Hawaii's cultural politics.

The Mauna Kea site has been prized by astronomers since the first telescope was built there in 1968. Today a dozen observatories — operated by US, European, Canadian and Japanese astronomers — dot the stark landscape near the summit.

But in recent years a growing Hawaiian movement has objected to the telescopes on the grounds that they desecrate the ancient volcano, according to local tradition the home of the snow goddess. Each new project has been hotly contested, and some Hawaiian groups have even called for existing telescopes to be pulled down.

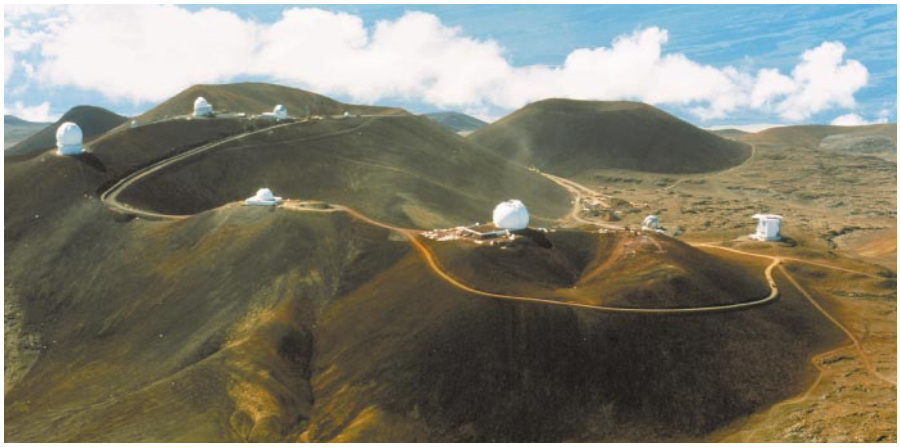
"The astronomy community was unprepared for this backlash," admits William Smith, president of the Association of Universities for Research in Astronomy (AURA), which manages the 8-metre Gemini telescope on Mauna Kea.

NASA began work in 1997 to link the two existing 10-metre Keck telescopes together as an optical interferometer to improve their ability to see fine detail, and to search for planets around other stars. But this means that at least four small 'outrigger' telescopes must be built near the large Keck domes.

A master plan for the Mauna Kea Science Reserve, completed last summer after two years of intense discussion and compromise among scientists, environmentalists and native Hawaiian groups, identified the interferometer as one of two approved expansion projects on the mountain.

Construction of the outriggers was to have begun in October, with a view to having the interferometer in use by next spring. But NASA is still embroiled in a lengthy process to ensure compliance with federal environmental and historic preservation laws — and ultimately to protect itself against lawsuits from groups opposed to astronomy on the mountain. Construction permits have not yet been granted for the outriggers.

In December, NASA published a draft environmental assessment of the project, which includes, among other provisions, a mitigation strategy for the wekiu bug (*Nysius wekiuicola*) that lives on Mauna Kea and is a candidate for protection under the Endangered Species Act. NASA hopes the assessment will satisfy environmental concerns: a more detailed environmental impact statement would add significantly to costs and delays.



ROYAL OBSERVATORY, EDINBURGH/SPL

Heavenly view: but planned new facilities have sparked environmental and cultural opposition.

Another consideration is the National Historic Preservation Act, which protects sites of cultural interest. In 1999, Hawaii's office of historic preservation concluded that the mountain meets the criteria for listing as a historic site. NASA has been discussing preservation and outreach measures with historic and cultural groups ever since.

Among the possibilities being discussed are having an archaeologist present while the outriggers are built, contributing to native Hawaiian schools, educating astronomers about the sacred traditions of Mauna Kea, and increasing the number of native-born Hawaiians working at the observatories.

"We want to do what is right," says John Lee, programme manager for the Keck Interferometer at NASA headquarters in Washington. "We don't want to poison the well" for astronomy on the mountain, he says. Most observatories on Mauna Kea will eventually need to modernize their telescopes.

Three new projects are also included in the master plan, including the ambitious Giant Segmented-Mirror Telescope, which is still in the conceptual stage, but at 30 metres diameter would be by far the largest optical telescope ever built.

NASA managers are aware of the pressure to resolve the dispute. Astronomers "see themselves losing their best Northern Hemisphere site", says Anne Kinney, science director for the agency's Origins programme. Further delays "would hold up every single project on the mountain". ■

Weapons lab seeks Mexican link

Rex Dalton, San Diego

Sandia National Laboratory, a nuclear weapons lab in New Mexico, has launched plans for an applied-technology laboratory on the border between the United States and Mexico.

Scientists and engineers at the Bi-National Sustainability Laboratory would work on developing technologies to make new products in the fields of energy, water, air quality, health care and information services.

The new lab would be sited in the border cities of Santa Teresa, New Mexico, and San Jeronimo, Chihuahua, Mexico. These are part of a jointly planned development area that includes a border crossing.

The lab is being promoted by Sandia's Advanced Concepts Group, which pursues new research ideas. "The idea is a mix of a scientific incubator, think-tank and technology start-up," says Sandia engineer Vipin Gupta, a member of the group.

Sandia officials hope the new lab will be funded equally by the US and Mexican governments, and be staffed by scientists from both countries. Sandia planners were in Mexico last week looking for an agency there to be the primary partner. The United States–Mexico Foundation for Science, a think-tank based in Mexico City, Mexico, is helping with project planning.

The cost of building and operating the lab has yet to be determined, although Sandia officials have employed an architect to create a model of the facility.

The location and focus of the lab are the idea of Sandia engineer Gerry Yonas. "We see this as a tremendous opportunity," he says. "But if we are going to make this concept a reality, we need a full partner in Mexico." ■

© 2001 Macmillan Magazines Ltd

World's largest research drilling ship under construction

Tokyo Construction of the world's largest and deepest drilling research vessel officially got under way this week in Okayama, Japan.

The vessel will be used for research projects managed by the international Ocean Drilling Program (ODP). It will be able to drill to a depth of 7 kilometres, more than three times the reach of the current ODP vessel. The $500 million project is due to be completed in 2004, with research use beginning in 2006.

ODP

Digging deep: the Ocean Drilling Program's vessel will drill to a depth of seven kilometres.

The ship's future had been in doubt after disagreements over the quality of research produced by similar projects (see *Nature* **388,** 409; 1997). Potential research projects include studying the movement of heat and rocks around seismogenic areas — where earthquakes originate — and searching for bacteria that survive below the Earth's surface.

http://www.jamstec.go.jp/jamstec-e/odinfo

Agent Orange may lead to child leukaemia

Washington The use of the defoliant known as Agent Orange by the United States in the Vietnam War may be linked to a form of leukaemia among the children of Vietnam veterans, a US Institute of Medicine (IOM) committee said last week.

Two studies found that the parents of children with acute myelogenous leukaemia, which represents 8% of childhood cancers, were more likely to have served in the war or used pesticides in the workplace. Another study revealed a high prevalence of the disease among children of Australian Vietnam veterans. Earlier reports had found insufficient evidence for the link.

The US military sprayed millions of litres of Agent Orange and other dioxin-containing herbicides on Vietnam and Cambodia between 1962 and 1971. The IOM has assessed the effect of Agent Orange on the health of veterans every two years since 1994.

Reactor needed to avert 'neutron drought'

Munich The International Group on Research Reactors has called on the German government to approve the operation of FRM-II, a new 20-megawatt neutron source located at Garching, near Munich. Scientists at the group's meeting last week in Munich warned that they will face a 'neutron drought' if the reactor is not approved.

FRM-II has been dogged by controversy since its conception. Unhappy with plans that FRM-II will use weapons-grade uranium, the federal government ruled in 1999 that the reactor should convert to low-grade uranium as soon as technically possible (see *Nature* **400,** 7; 1999). It has also recommended that the state government in Bavaria, where FRM-II is based, build its own storage facility for the fuel.

Bleak outlook for funding of stem-cell research

Washington The chances that the US government will fund medical research involving human embryos receded further

© 2001 Macmillan Magazines Ltd

last week when the Department of Health and Human Services indefinitely postponed a review of proposals for stem-cell research.

The National Institutes of Health (NIH) published guidelines last year stating that only embryos left over from fertility treatments should be used in federally funded research. The Human Pluripotent Stem Cell Review Group was created to oversee adherence to the guidelines by NIH-funded labs, but the group's meeting planned for this week has now been called off.

The Department of Health and Human Services has said it needs to review the NIH guidelines before the group can meet. President George W. Bush has already said that he opposes the use of public funds for any research that destroys human embryos.

India paves way to launch capability

New Delhi Indian scientists restored some pride to their space programme last week by successfully launching a rocket that may one day be capable of putting communications satellites into geostationary orbits. A rocket failure had ended the original launch last month of the 14 billion rupee (US$300 million) Geosynchronous Satellite Launch Vehicle (GSLV) (see *Nature* **410,** 619; 2001).

India currently uses Europe's Ariane rockets to launch domestic communications satellites. The GSLV will allow India to end this dependence and to enter the global satellite-launching business. The rocket's manufacturer, the Indian Space Research Organization, now has to increase its payload capacity to allow commercial use.

Lift off: India hopes to see space profits soar.

ISRO

Call for stronger regulation of research on humans

Washington **A voluntary accreditation system for all research involving human participants is needed to protect the safety of human subjects, a US Institute of Medicine (IOM) report said last week.**

The current regulatory system relies on several institutional review boards which, the IOM says, have not adapted to the rapid expansion of clinical research. The report calls for an independent body to develop standards based on federal regulations and on the accreditation programme currently being developed for US Department of Veteran Affairs medical centres. Accreditation should apply to human research at institutions in all fields, including social and behavioural sciences, it says.

The report follows the death in 1999 of a volunteer in a gene-therapy trial and the subsequent suspension of several clinical research programmes.

http://www.nap.edu/books/0309073286/html

Waste mix-up creates radioactive fertilizer

London Farmers in nuclear-free Norway have been spreading low-level radioactive waste on their land for nine years after a pipeline mix-up, it was revealed last week.

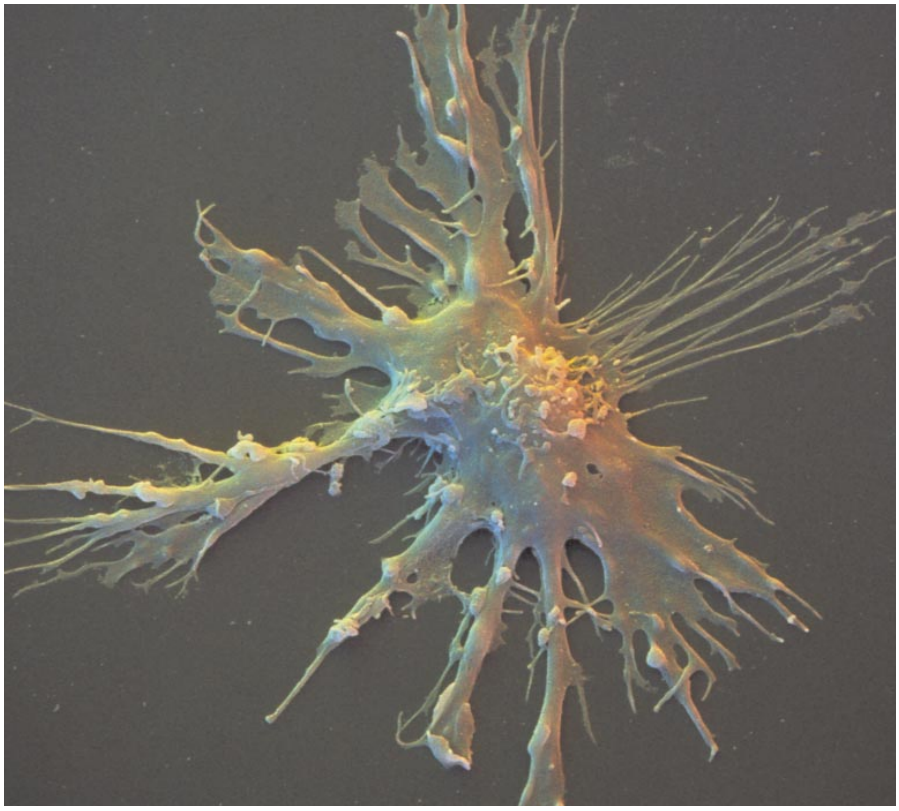
The country's Institute for Energy Technology (IFE) has admitted that, between 1991 and 1999, cooling water from a research reactor that should have been piped out to sea was instead directed into sewers beneath the city of Halden in southeast Norway.

Some of the resulting sewage sludge containing the nuclear waste was sold to farmers for use as fertilizer. An IFE spokesperson insisted that there was no risk to health, as radioactive emissions from the waste water were well within safety limits. The pipeline has now been correctly connected.

© 2001 Macmillan Magazines Ltd

Cinderella goes to the ball

After decades of neglect, research into 'innate' immunity has moved to centre stage. Phyllida Brown explores the excitement that now surrounds an evolutionarily ancient arm of our immune system.



SPL

Trigger happy: receptors on dendritic cells help to fire up the innate immune response.

The innate immune response — an inflammatory reaction that helps control infections in their early stages — has long been the Cinderella of immunology. For decades, innate immunity was dismissed as an evolutionary throwback, providing no more than a quick-and-dirty holding operation before giving way to the exquisitely complex adaptive immune response. Studies into how the adaptive immune system churns out cells and antibodies tailored to recognize particular pathogens attracted all the glamour.

But now all that is changing. Over the past few years, researchers have begun to realize that the innate immune response is a powerful screening tool. In the early hours of an infection, it can distinguish between different classes of pathogenic bacteria, viruses and fungi. And it is becoming clear that the innate immune response is also crucial for prodding the slower-acting adaptive system into action. "The field has suddenly heated up," says Alan Aderem, an immunologist at the Institute for Systems Biology in Seattle.

It is innate immunity that causes the redness and swelling that soon erupts around an infected wound, and the initial aches and fever that precede a bout of flu. Such reactions help to prevent invading pathogens from spreading rapidly around the body — which is important, because the adaptive immune system might not kick in for several days.

The chief cause of the fresh excitement is a rapidly accumulating body of research into a family of molecules called Toll-like receptors (TLRs), which have a central role in triggering innate immune responses. These receptors sit on the surface of two classes of mammalian white blood cells — macrophages and dendritic cells — and on the cells of certain other tissues such as the epithelia that line our guts and nasal passages.

The TLRs owe their name to a closely related receptor called Toll, first identified in the fruitfly *Drosophila*. Invertebrates lack an adaptive immune system, but when Toll and its allies are stimulated by the presence of foreign microorganisms, they initiate the production of antimicrobial compounds that target the invaders. Mammalian innate immunity works differently, producing molecules called cytokines that cause inflammation. But the fact that the TLRs are so similar to Toll reveals that they are part of an ancient defence mechanism that has been conserved through evolution.

Screen test

Researchers are now discovering that each type of TLR recognizes specific molecular patterns carried by different pathogens. These pathogen-associated molecular patterns, or PAMPs, are usually embedded in structures that are essential to their bearers' survival. As such, they evolve slowly and tend to be shared by whole classes of bacteria, viruses or fungi. And, as different classes of microbe carry different PAMPs, a macrophage or dendritic cell can use its family of TLRs to classify the invader, and respond appropriately. "What we are starting to see is a digital information system, rather like barcode recognition," says Aderem.

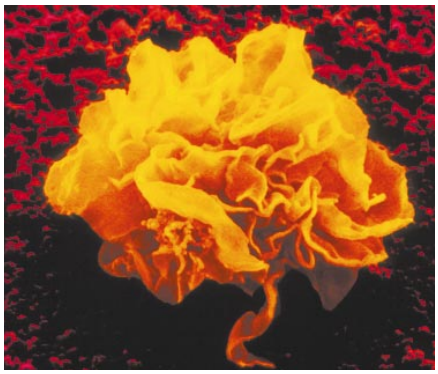
Since the first mammalian TLR, now called TLR4, was identified in 1997 by Charles Janeway and Ruslan Medzhitov at Yale University in New Haven, Connecticut[1], nine more have been added to the list[2–5]. Researchers have also made rapid progress in linking individual TLRs with particular PAMPs. TLR4 is activated by lipopolysaccharide (LPS), a component of the cell wall of Gram-negative bacteria such as *Salmonella*[6]. It has since been shown also to recognize other bacterial PAMPs[7], and an important viral protein[8]. TLR2, meanwhile, responds to a glycolipid called lipoarabinomannan[9,10], a component of the cell walls of mycobacteria, a group that includes the tuberculosis pathogen. It is also activated by the LPS of *Leptospira* bacteria[11]. And in this issue of *Nature*[12], Aderem's team shows that TLR5 is activated by flagellin, a protein found on the surface of flagellate bacteria such as *Listeria*.

With some TLRs responding to more than one PAMP, and other TLRs operating together, recognizing combinations of PAMPs specific to particular classes of pathogen[13], Medzhitov argues that the 10 known receptors can deal with around 20 specific PAMP combinations. "That's enough to recognize practically any class of infective agent," he says.

Once activated by a PAMP, a TLR triggers a cascade of cellular signals. Most researchers

© 2001 Macmillan Magazines Ltd

SPL



On the defensive: the 'Toll-like' receptors on cells such as macrophages (bottom left) can recognize invading pathogens such as *Salmonella* (right) and kick-start an immune response. They derive their unusual name from similar proteins found in fruitflies (top).

have so far focused on signalling pathways that eventually activate a transcription factor called NF-κB, which regulates the activity of the genes that produce cytokines such as tumour necrosis factor-α and interleukin-1. But there is evidence that other signalling pathways are also switched on when TLRs are activated[14]. In general, the cytokines make blood vessels leakier at the site of infection, allowing fluid and proteins to pass into the tissues. Cytokines also attract other white blood cells to the scene.

Concerted effort

Exactly how PAMPs activate their corresponding TLRs remains largely mysterious, and it seems likely that other 'helper' molecules are involved. For example, on the surface of Gram-negative bacteria, LPS is complexed with another molecule called LPS-binding protein (LBP). This complex is recognized by a molecule called CD14, which is carried by macrophages[15]. Immunologists believe that it is probably the trio of LPS, LBP and CD14 that activates TLR4 (ref. 14).

Although hard evidence is still scant, most researchers suspect that different TLRs trigger subtly different combinations of cytokines, each appropriate for different classes of infectious agent. For example, one might expect that TLRs that recognize viral proteins would trigger the release of cytokines that instruct other cells to make potent antiviral compounds such as interferons. But the overlapping roles of different signalling pathways and cytokines make it difficult to turn this speculation into proven fact.

The buzz surrounding TLRs is not just a result of their role in triggering innate immune responses, however. There is also growing evidence that activated TLRs send vital wake-up signals to the adaptive immune system. "Until TLRs were discovered, the innate immune system was considered to be just an evolutionary rudiment," says Medzhitov. "But we now know that it plays a key role in adaptive immunity."

> We are starting to see a digital information system, rather like barcode recognition.

Receptive audience: Charles Janeway (left) and Ruslan Medzhitov were the first to identify Toll-like receptors in mammalian cells.

ELIZABETH KOPP

Janeway has long championed this idea. Back in 1989, when he first suggested that the innate immune response is triggered by microbial molecules such as LPS, he also argued that the same signals trigger the innate system to activate 'naïve' cells of the adaptive system and turn them into armed and proliferating attackers[16]. Janeway based his argument on studies of adjuvants, components that must be added to most vaccines to ensure that they provoke a sufficiently vigorous immune response. Many adjuvants contain bacterial molecules, and Janeway argued that these are recognized by the innate immune system, which prods the adaptive system into action.

T cells are key players in the adaptive immune system. Different T cells carry different receptors that recognize distinctive molecules known as antigens. At any one time, the immune system includes a vast array of circulating T cells with different specificities, each present in small numbers. When the body comes under attack, the relevant T cells are activated and proliferate rapidly. (Similar processes underlie the production of specific antibodies by B cells.)

For whom the cell tolls

The key to effective adaptive immunity lies in ensuring that this activation only occurs in response to antigens that present a threat — reactions to most of the body's own proteins, or innocuous antigens such as those in our food, should be avoided. Immunologists have long known that macrophages and dendritic cells are involved in processing antigens so that they can be recognized by T cells. But these antigen-presenting cells only activate T cells when they also produce 'co-stimulatory' molecules.

The fact that macrophages and dendritic cells are also involved in innate immunity might have suggested a link between the two arms of the immune system. But supporting evidence for Janeway's hypothesis that innate immune responses help trigger adaptive immunity has been boosted by studies of the TLRs. The best known of the co-stimulatory molecules belong to a group of proteins called B7, produced by macrophages and dendritic cells. When Janeway and his colleagues discovered TLR4 in 1997, they showed that its activation causes the cells involved to produce B7 in addition to cytokines[1].

Then, last December, researchers led by Shizuo Akira of Osaka University in Japan added another line of evidence. They observed that sequences found mainly in bacterial DNA and rarely in the DNA of higher organisms — called CpG repeats — are recognized by TLR9 (ref. 17). As CpG repeats in vaccines that are based on naked DNA stimulate the adaptive immune system to respond to these vaccines, this strength-

© 2001 Macmillan Magazines Ltd

ened the link between the two arms of the immune system.

The latest twist in the tale is that the TLRs might also help to explain a problem that has long puzzled immunologists: why innate and adaptive immune responses are sometimes triggered by the body's own proteins in the absence of any infection. For example, when tissue is damaged in major surgery, the area around the incision sees soaring levels of cytokines and inflammation, however sterile the conditions. This inflammatory response can also trigger a transient burst of antibodies against the body's own tissues.

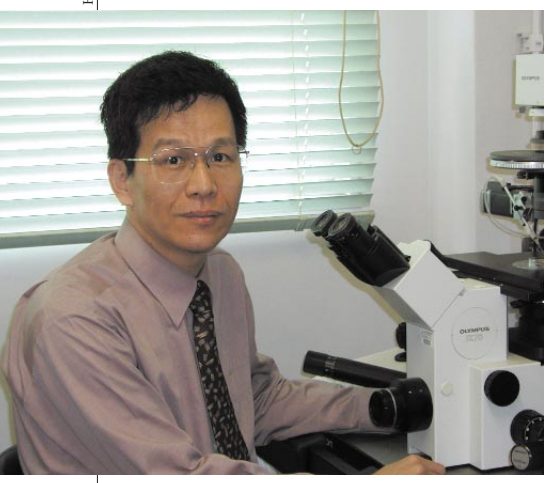
Danger signs

In 1994, Polly Matzinger of the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, came up with a provocative suggestion[18]. She said that it was simplistic to argue that the immune system distinguishes between 'self' and 'foreign' antigens. Instead, Matzinger proposed that the key to triggering immunity is an association between an antigen and a 'danger' signal. In addition to molecules produced by pathogens, she said that the signals could include products of damaged tissue, such as proteins that only emerge from cells when their membranes are ruptured. In cellular terms, this is like entering a building and seeing blood spattered on the walls.

Evidence is now emerging that some TLRs may indeed be involved in responding to such danger signals. For example, last month, researchers at the University of Pennsylvania Medical Center in Philadelphia reported that fibronectin, a molecule released by damaged tissue, can activate TLR4 (ref. 19). This same receptor is activated by heat-shock proteins, which are produced by cells under stress[20]. Matzinger is delighted: "This is exactly what I have been saying."

As our understanding of the TLRs deepens, immunologists are starting to look for variation in the genes encoding TLRs and their expression to see whether this affects our susceptibility to disease. "We have known for a long time that different individuals have different responses to bacteria, for example, and this could be one of the reasons," says Marco Colonna of the Basel Institute for Immunology in Switzerland. Some people, for instance, are particularly prone to *Staphylococcus* infections, and Aderem suggests that deficiencies in the production of certain TLRs, or the signalling pathways that lie downstream of the receptors, might explain why. If so, he says, it might be possible to devise therapies to circumvent the blocked signalling pathways.

HOSHINO KATSUAKI



Missing links: Shizuo Akira has reinforced the suggestion that the two arms of the immune system are connected.

The innate immune response will never again be seen as a crude leftover from our evolutionary past.

Equally important, says Aderem, the receptors might also be involved in inflammatory diseases such as atherosclerosis, which is now widely believed to involve chronically activated macrophages. If certain TLR-related signalling pathways could be blocked without compromising vulnerability to infection, it might be possible, he argues, to design drugs that prevent the inflammatory process from running awry.

A rousing chorus

But there is one caveat. The TLRs are not the only recognition tool in the armoury of the innate immune system, nor are they its only way of waking up the adaptive immune system. Mammalian immune defences are famously rich in cells and cytokines with overlapping functions, and back-up signalling pathways. And evidence is already beginning to emerge for such secondary mechanisms that operate alongside the TLRs. In this issue of *Nature*[21], Colonna and his colleagues report on experiments using mice in which they examined a different receptor, called TREM-1, which is carried by other types of white blood cells such as neutrophils.

TREM-1 is activated by exposure to bacteria such as *Pseudomonas aeruginosa* or *Staphylococcus aureus*, although the precise molecular pattern it responds to is unknown. Colonna and his colleagues have now found that the receptor can start a signalling cascade that results in inflammation through a pathway that does not involve NF-κB. They have also found that TREM-1 may play a key role in septic shock, a systemic reaction that develops if massive quantities of inflammatory cytokines are released in response to a heavy load of microorganisms, for example in peritonitis.

Cool it: Alan Aderem thinks that studies of Toll-like receptors will help to design treatments for inflammatory disease.

More evidence for the diversity of signalling pathways involved in innate immunity has come from Akira's group. Until recently, a protein called MyD88 was thought to be key to TLR-triggered signalling. But new research from Akira's team indicates that dendritic cells from mice lacking MyD88 can still activate T cells, suggesting that this aspect of TLR4-triggered signalling uses an alternative pathway[22].

Clearly, there are plenty of unanswered questions. But as researchers scramble to uncover the workings of the TLRs and innate immunity, they may hit upon new strategies for the development of drugs and vaccines. One thing is for sure: the innate immune response will never again be regarded as a crude leftover from our evolutionary past. For Cinderella, the ball is far from over. ■

Phyllida Brown is a science writer based in Exeter.

1. Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A. Jr *Nature* **388,** 394–397 (1997).
2. Rock, F. L., Hardiman, G., Timans, J. C., Kastelein, R.A. & Bazan, J. F. *Proc. Natl Acad. Sci. USA* **95,** 588–593 (1998).
3. Chaudhary, P. M. *et al. Blood* **91,** 4020–4027 (1998).
4. Takeuchi, O. *et al. Gene* **231,** 59–65 (1999).
5. Chuang, T. & Ulevitch, R. J. *Biochim. Biophys. Acta* **1518,** 157–161 (2001).
6. Poltorak, A. *et al. Science* **282,** 2085–2088 (1998).
7. Lien, E. *et al. J. Biol. Chem.* **274,** 33419–33425 (1999).
8. Kurt-Jones, E. A. *et al. Nature Immunol.* **1,** 398–401 (2000).
9. Means, T. K. *et al. J. Immunol.* **163,** 3920–3927 (1999).
10. Underhill, D. M., Ozhinsky, A., Smith, K. D. & Aderem, A. *Proc. Natl Acad. Sci. USA* **96,** 14459–14463 (1999).
11. Werts, C. *et al. Nature Immunol.* **2,** 346–352 (2001).
12. Hayashi, F. *et al. Nature* **410,** 1099–1103 (2001).
13. Ozinsky, A. *et al. Proc. Natl Acad. Sci. USA* **97,** 13766–13771 (2000).
14. Aderem, A. & Ulevitch, R. J. *Nature* **406,** 782–787 (2000).
15. Tobias, P. S., Soldau, K., Gegner, J. A., Mintz, D. & Ulevitch, R. J. *J. Biol. Chem.* **270,** 10482–10488 (1995).
16. Janeway, C. A. Jr *Cold Spring Harbor Symp. Quant. Biol.* **54,** 1–13 (1989).
17. Hemmi, H. *et al. Nature* **408,** 740–745 (2000).
18. Matzinger, P. *Annu. Rev. Immunol.* **12,** 991–1045 (1994).
19. Okamura, Y. *et al. J. Biol. Chem.* **276,** 10229–10233 (2001).
20. Ohashi, K., Burkart V., Flohe, S. & Kolb, H. *J. Immunol.* **164,** 2064–2069 (2000).
21. Bouchon, A., Facchetti, F., Weigand, M. A. & Colonna, M. *Nature* **410,** 1103–1107 (2001).
22. Kaisho, T., Takeuchi, O., Kawai, T., Katsuaki, H. & Akira, S. *J. Immunol.* **166,** 5688–5694 (2001).

© 2001 Macmillan Magazines Ltd

Even 'free access' is still beyond the means of most scholars in Africa

Sir— The draft sequence of the human genome (*Nature* **409,** 814–958; 2001) is a landmark achievement that will satisfy intellectual curiosity and biotechnological ambitions for a considerable time. *Nature* is to be applauded for its open publication of the sequence and the many excellent research articles and analyses that accompany it. However, although freely available, access is far from free for all.

Here in west and central Africa we inhabit a communication wasteland, with only fragmentary evidence of the electronic capacity of the twenty-first century. Faced with computer scarcity, limited bandwidth connectivity to the Internet, and few skilled educators to explain molecular biology, access to the human genome sequence is not free. Journal availability throughout the whole region is virtually non-existent, and the costs of personal subscriptions prohibitive. In contrast, all editions of *Nature* and many other journals are indeed freely available to most students and researchers in Europe and North America, through institutional subscriptions.

Africa came away from the green revolution empty handed; the biotechnology revolution has all but passed it by. If Africa is to exploit the new-found knowledge of the human genome and to participate effectively in the biotechnology revolution, the bottlenecks must be removed.

One critical constraint is access to information. With modest donor support and the goodwill of publishers, online access to journals, books and patents could inspire students and teachers with the latest scientific discoveries, rejuvenate the aspirations of researchers, and provide policy-makers with insights into issues such as biosafety, intellectual property and commercial opportunities. Online information could provide a keystone for Africans to develop innovations that tackle African needs in a continent ravaged by HIV, malaria, malnutrition and poverty.

Paul Keese
International Institute of Tropical Agriculture, Oyo Road, Ibadan, Nigeria

Precarious life in Spain

Sir— As you have frequently reported, Spanish scientists face many obstacles in their funding and careers (see, for example, *Nature* **408,** 397 & **405,** 723; 2000). Young Spanish researchers lack employment rights such as affiliation to the social security system or health benefits. They depend on grants and are considered by the government as students, yet they have to pay tax and must engage exclusively in research and lecturing.

Pre- and postdoctoral scientists are usually not represented in the elected body of their centres. Further, because of the pyramidal structure of the Spanish career system, young researchers have fewer opportunities as they become more senior. Often their careers are interrupted while awaiting the next grant, yet they receive no unemployment benefits. The situation is even worse in light of Spain's small research budget, few new permanent positions and nepotism in some research centres.

In February, more than 4,000 young researchers denounced this situation in front of the Ministry for Science and Technology in a demonstration organized by a group called Precarios: the Young Researchers' Federation. Their name recalls both the Spanish word for people on a grant, *becarios*, and their precarious situation (see http://www.precarios.org). They presented a manifesto signed by some 3,000 senior Spanish scientists, demanding a new status for young researchers, including social benefits and an increase in investment in R&D budgets.

Proper recognition in the early stages of scientific careers is a necessary step towards the inclusion of Spain in the group of scientifically most advanced countries.

Manuel J. Pérez Mendoza, for Precarios
Grupo de Investigación en Carbones, Departamento de Química Inorgánica, Universidad de Granada, Av. Fuentenueva s/n., 18071 Granada, Spain

Let's reward innovative action against poverty

Sir— Each year, the world rightly honours the best contributors to science, economy, literature and peace. However, these accolades stop short of their ultimate objective, which is the improvement of the human condition, especially where living conditions are atrocious. There is no commensurate recognition for scientific work that directly improves the quality of life of the poor and underprivileged.

It should not be difficult to design such a prize, because there are now indicators to measure socioeconomic development. What has emerged is that science, technology, creative economic principles, elegance in literature and even sacrifices for peace do not in themselves reduce poverty and alienation. Progress requires deliberate attempts to link these advances with efforts to provide basic needs, empower the poor and banish inequity and injustice.

What should be singled out for support and commendation are ways to bring together knowledge and grass-roots community action. Underdevelopment is the greatest threat to world peace and global environmental security, and therefore deserves special intellectual and scientific consideration. A serious, well-publicized prize to honour those who have used imagination and innovation to push back disease, deprivation and injustice is the most important prize of all, if the objective is survival of a civilized planet.

Arnoldo Ventura
Office of the Prime Minister, 1 Devon Road, Kingston 6, Jamaica, West Indies

Entropy illustrates the flexibility of Chinese

Sir— The Words essay by Alan L. Mackay (*Nature* **410,** 19; 2001) illustrates the challenges that the Chinese language faces in the new millennium. As a native speaker, I would like to offer a different, perhaps more accurate interpretation of the etymology of the character 熵 (*shang*), which means 'entropy'.

The translation of entropy into 熵 is a perfect example of the adaptive nature of the Chinese language. It reflects the equation $S = H/T$, where S is entropy, H is enthalpy (焓— *han*), and T is the absolute temperature. The left half, 火 (fire), hints that the character 熵 has something to do with heat or temperature, and the right half, 商, means 'quotient' instead of Mackay's interpretation of 'merchant'.

This difference is understandable, as an English speaker who knows the Greek word-root *trope*, which denotes turning and changing, would naturally associate the character 商 with 'merchant'. But an educated Chinese speaker does not readily associate a scientific character such as 熵 with a merchant. In the traditional Chinese value system, merchants were considered not quite highbrow. The association with quotient is much more spontaneous, as a difficult character such as 熵 confers an implicit sense of being about learned matter.

The translation of enthalpy to 焓 is equally exquisite. The left part of the character means 'fire' as in entropy; the right part, 含 means 'content'.

These are just two examples of how new Chinese characters were coined as the need arose. None of the new characters that I know of was created *de novo*, in keeping both with the long tradition of the language, and with its creativeness and adaptability.

Jian Feng (冯 简)
Department of Physiology and Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, New York 14214, USA

© 2001 Macmillan Magazines Ltd

Access all areas?

The Internet has revolutionized how we live and work — with dynamic new technologies it is set to change the way in which science reaches its audience.

This week's Commentary pages offer three highly individual perspectives on future developments in access to the primary scientific literature. On this page, Tim Berners-Lee and James Hendler present their vision of a semantic web, where new technologies will allow computers, as well as people, to understand and communicate with each other and hence to revolutionize scientific publishing. On page 1024, Stevan Harnad proposes "freeing" the refereed research literature by implementing an authors' searchable, self-archiving system that would turn publishers into providers of a peer-reviewing service rather than producers of journals. And on page 1026, Ira Mellman argues that the "public library of science" initiative should focus on making journals' content free-access six months after publication, rather than on pressurizing them to permit authors to re-display published material on other sites.

These articles are also published, in slightly different form, on *Nature*'s website as part of its current web debate on electronic publishing initiatives in science (see *Nature* **410,** 613; 2001). Readers wishing to participate in the debate are invited to view http://www.nature.com/nature/debates/e-access. Contributions can also be submitted to Correspondence via corres@nature.com. In either case, publication will be offered according to the criteria described on page 613 of the 5 April issue.

Publishing on the semantic web

The coming Internet revolution will profoundly affect scientific information.

Tim Berners-Lee and James Hendler

To predict the future of scientific publishing on the World-Wide Web, it is important to understand how web technology is changing. We are in the early days of a new web revolution, one that will have profound implications on web publishing, and on the nature of the web itself. Just as current web technology is changing the world of publishing, the new semantic web technology (http://www.w3.org/2001/sw) may change the way scientific knowledge is produced and shared.

The web was designed as an information space, with the goal not only that it should be useful for human–human communication, but also that machines would be able to participate and help users communicate with each other. A major obstacle to this goal is the fact that most information on the web is designed solely for human consumption. Computers are better at handling carefully structured and well-designed data, yet even where information is derived from a database with well-defined meanings, the implications of those data are not evident to a robot browsing the web. More information needs to be in a form that the machine can 'understand' rather than simply display.

The concept of machine-understandable documents does not imply some magical artificial intelligence allowing machines to comprehend human mumblings. It relies solely on the machine's ability to solve well-defined problems by performing well-defined operations on well-defined data. So, instead of asking machines to understand people's language, the new technology, like the old, involves asking people to make some extra effort, in repayment for which they get major new functionality — just as the extra effort of producing HTML mark-up is outweighed by the benefit of having content searchable on the web.

A new set of languages is now being developed to make more web content accessible to machines. The Semantic Web Activity, run by the World Wide Web consortium, is defining new web technologies that will enable successively better tools that make it easier for people to create machine-readable content and make it widely available.

What impact might this have on scientific publishing? In the next few years, we expect that tools for publishing papers on the web will automatically help users to include more of this machine-readable mark-up in the papers they produce. Where a current tool using XML (see http://www.nature.com/nature/webmatters/xml/xml.html) can allow a user to assert that some part of a document is about an 'experiment', the new languages will let the scientist express that the experiment uses certain chemicals and reagents; that the system used involved some particular organic matter; that the experiment produced gels with certain DNA information on them (and that the images of these gels are located in particular places on the web); and so on.

Papers that include this new mark-up language will be found by new and better search engines, so users will be able to issue significantly more precise queries. More importantly, experimental results can themselves be published on the web, outside the context of a research paper. So a scientist can design and run an experiment, and create an emerging web page containing the information that he or she wants to share with trusted colleagues (see Fig. 1). Finding out about experiments and studies in progress will be easy, and work can be modified as a result of interaction with peers, with less need to wait for formal publication. Just as preprints challenge established journals' online versions, these new 'papers in progress' will be a significant challenge to online scientific publishers.

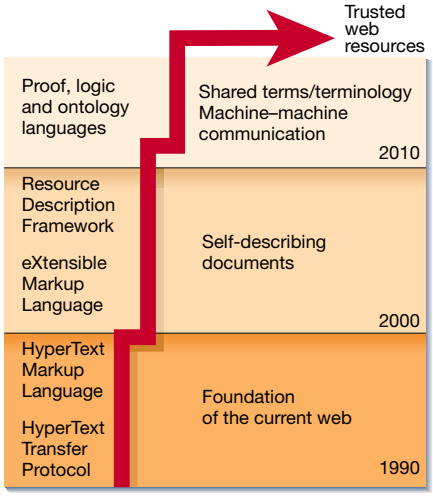


Figure 1 **Layers of the semantic web are built as new languages and tools anchored in XML. They build towards a world of trusted information shared among collaborating groups of users.**

In the long run, the effects on publishing may be far more profound. There is an eternal conflict between operating rapidly as a small group and taking the time to communicate more widely. The former is more efficient but produces a subculture whose concepts and results are not understood by others. The latter can be painfully slow. The world works as a spectrum between these extremes, with a tendency to start small — from the personal idea — and filter over time towards a wider commonality of concept. The joining together of subcultures when there is a need for a wider common language is an essential process in the

© 2001 Macmillan Magazines Ltd

development of human communication.

The semantic web will facilitate the development of automated methods for helping users to understand the content produced by those in other scientific disciplines. On the semantic web, one will be able to produce machine-readable content that will provide, say, automated translation between the output of a scientific device and the input of a datamining package used in some other discipline, or a self-evolving translator that allows one group of scientists directly to interact with the technical data produced by another.

These new products will help users communicate with each other even before a common vocabulary has developed to express the concepts they have in common. The semantic web will provide unifying underlying technologies to allow these concepts to be progressively linked into a universal web of knowledge. Thus it will aid in breaking down the walls of miscommunication, allowing researchers to find and understand products from other scientific disciplines. The very notion of a journal of medicine separate from a journal of bioinformatics, separate from the writings of physicists, chemists, psychologists and even kindergarten teachers, will some day seem as out of date as the print journal is becoming to our graduate students.

Talking point: Tim Berners-Lee backs more participation from computers in web publishing.

Does this sound like a crazy science-fiction dream? A decade ago, who would have believed a web of text, conveyed by computer, would challenge a 200-year-old tradition of academic publishing? ■

Tim Berners-Lee, inventor of the World-Wide Web, is at the World Wide Web Consortium, MIT, 545 Technology Square, Room NE43-356, Cambridge, Massachusetts 02139, USA (http://www.w3.org/People/Berners-Lee). James Hendler is in the Computer Science Department, University of Maryland, College Park, Maryland 20853, USA.

The self-archiving initiative

Freeing the refereed research literature online.

Stevan Harnad

Unlike the authors of books and magazine articles, who write for royalty or fees, the authors of refereed journal articles write only for 'research impact'. To be cited and built on in the research of others, their findings have to be accessible to their potential users. From the authors' viewpoint, toll-gating access to their findings is as counterproductive as toll-gating access to commercial advertisements.

With the online age, it has at last become possible to free the literature from this unwelcome impediment. Authors need only deposit their refereed articles in 'eprint' archives at their own institutions; these interoperable archives can then all be harvested into a global virtual archive, its full contents freely searchable and accessible online by everyone (see Box).

Unlike the royalty/fee-based literature, which constitutes the vast majority of the printed word, the special, tiny literature of refereed journal articles is, and always has been, an 'author giveaway'. Researchers never benefited from the fact that people had to pay access tolls to read their papers (as subscriptions, and for the online version, site-licences or pay-per-view). On the contrary, those access barriers represent impact barriers for researchers, whose careers and standing depend largely on the visibility and uptake of their research.

There are currently at least 20,000 refereed journals across all fields of scholarship, publishing more than 2 million refereed articles each year. The amount collectively paid by those of the world's institutions which can afford the tolls for just one of those refereed papers averages $2,000 per paper[1]. In exchange for that fee, that particular paper is accessible to readers at those, and only those, paying institutions.

The research libraries of the world can be divided into the (minority) Harvards and the (majority) Have-nots — the last by no means limited to the developing world. It is obvious how the Have-nots would benefit from free access to the entire refereed literature, for without it their meagre serials budgets can afford only a pitifully small portion. But not even Harvard can afford access to anywhere near all of the literature (see http://fisher.lib.virginia.edu/newarl/index.html). Hence, most refereed articles are inaccessible to most researchers. For the authors, this means that much of their potential impact is lost. And it is solely this curtailed research impact and access that is being purchased by the collective $2,000 outlay per article mentioned above.

This is the way things had to be in the past, when publishing as print-on-paper was the only medium, and the sizeable costs of printing and distribution had to be recovered somehow. The new online era may be threatening the majority, royalty/fee-based literature (books, magazine articles) in the form of digital piracy; but for the 'giveaway' research literature, it has at last made it possible to eliminate all those counterproductive access/impact barriers.

Not all costs have vanished, of course. Although the costs of printing and distribution (and their online successors, such as publishers' PDF page-images) are no longer essential ones, the cost of the quality-control and certification that differentiates the refereed literature from an unfiltered, anarchic vanity press still needs to be paid. Paper and PDF files have become mere options, purchasable by those who want and can afford them. Refereeing, however, is essential.

Essential costs of refereeing

Refereeing (peer review) is the system of evaluation and feedback by which expert researchers assure the quality of each others' research findings. Referees' services are donated free to virtually all scientific journals, but there is a real cost to implementing the refereeing procedures, which include archiving submitted papers on a website; selecting appropriate referees; tracking submissions through rounds of review and author revision; making editorial judgments, and so on.

The minimum cost of refereeing has been estimated as $500 per accepted article by the American Institute of Physics (see http://documents.cern.ch/archive/electronic/other/agenda/a01193/a01193s4t8/transparencies/Doyle.ppt), but that figure almost certainly has inessential costs wrapped into it (for example, the creation of the publisher's PDF). I think that the true figure for peer-review implementation alone across all refereed journals probably averages closer to $200 per article, or even lower. Hence, quality-control costs account for only 10% of the collective tolls actually being paid per article.

Can this situation, in which the authors' and referees' giveaways are needlessly being held hostage to obsolete printing costs and cost-recovery methods, be remedied? Note that it is not simply a matter of lowering the financial access barriers: even if those were slashed by 90%, most researchers would still be unable to access most research papers. There is an optimal solution, and it is inevitable: the refereed research literature must be freed online for everyone, everywhere, for ever. The irreducible 10% or so quality-control cost need no longer be paid for by readers' institutions; it can be paid in the form of quality-control service costs, per paper published, by authors' institutions, out of their savings on subscription costs.

Journal publishers certainly will not scale

© 2001 Macmillan Magazines Ltd

down to becoming only quality-control providers of their own accord. Nor can libraries effect such a transition on their own. And authors cannot and should not be expected to stop submitting their research to established high-quality, high-impact journals in preference for new, alternative journals just because those are prepared to provide stand-alone quality control right now. Journal niches are largely filled already, and immediate careers and standing are far more important to researchers than the potential long-term benefits of risky sacrifices.

But researchers can hasten the optimal and inevitable outcome without any sacrifice or risk. The entire refereed journal literature can be freed, virtually overnight, without authors having to give up their established refereed journals, by a method that a portion of the physics community has already shown to work. These physicists have since 1991 been publicly self-archiving their research papers online — both before and after refereeing (preprints and postprints) — in the physics 'eprint archive' at http://www.arxiv.org. This archive currently holds 150,000 articles. The number of new articles being self-archived there is currently about 30,000 annually, and increasing by some 3,500 papers each year. The archive, with its 14 mirror-sites worldwide, gets about 160,000 user hits each weekday at its US site alone. So there is no doubt that self-archiving is feasible, and that when papers are thus made freely accessible online, they are heavily used.

But although these physicists have shown the way to free the refereed research literature, authors in other disciplines have been slow to realize that the system can work for them too. They have assumed that there must be something unique about physics that makes self-archiving work. This misapprehension has been encouraged by the incorrect impression that the physics archive contains only unrefereed preprints, and that self-archiving somehow compromises the quality control of journals.

Yet absolutely nothing has changed in peer review in physics. The same authors who self-archive continue to submit all their papers to their journals of choice, just as they always did, and virtually all the papers in the archive appear in refereed journals about 12 months after journal submission. The only thing that has changed is that a growing portion of the refereed literature in physics is accessible, free for all, online. Yet even in physics, self-archiving is still growing far too slowly: at the present linear growth rate it will be another decade before the entire physics literature is online and free.

Institution-based self-archiving

There is now a way both to accelerate the rate of self-archiving in physics and to extend the practice to the other disciplines (see Box). My original 'subversive proposal'[2] to free the refereed literature through author self-archiving fell largely on deaf ears because self-archiving in an anonymous FTP archive or a web home page would be unsearchable, unnavigable, irretrievable and hence unusable. Nor has centralized archiving, even when made available to other disciplines, been catching on fast enough either (it has taken three years for the number of articles in http://cogprints. soton.ac.uk to reach 1,000).

Global goal: Stevan Harnad wants literature freed online for everyone, everywhere, for ever.

The new breakthrough is agreement on metadata tagging standards that make the contents of distributed archives interoperable, hence harvestable into one global virtual archive, all papers searchable and retrievable by everyone for free. The open archives initiative (OAI) at http://www.openarchives.org has now provided the metadata tagging standards and a registry for all OAI-compliant eprint archives. The self-archiving initiative at http://www.eprints.org is providing free software for institutions to create OAI-compliant archives, interoperable with all other open archives, ready to be registered and for their contents to be harvested into searchable global archives, interlinked to one another by citations (see http://cite-base.ecs.soton.ac.uk/cgi-bin/search).

Distributed, institution-based self-archiving benefits research institutions in three ways. First, it maximizes the visibility and impact of their own refereed research output. Second, by symmetry, it maximizes their researchers' access to the full refereed research output of all other institutions. Third, institutions themselves can hasten the transition to self-archiving and so more quickly reduce their library's annual serials expenditures to 10% (paid to journal publishers for refereeing their submissions).

The institutional library can help researchers to do self-archiving and can maintain the institution's own eprint archives as an outgoing refereed collection for external use, in place of the old incoming collection via subscription costs for internal use. Institutional library consortial power can also be used to provide leveraged support for journal publishers who commit themselves to a timetable of downsizing on the way to becoming pure quality-control service providers (http://ww.arl.org/sparc/home/index.asp). ■

Stevan Harnad is in the Intelligence/Agents/Multimedia Group, Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

1. Odlyzko, A. M. "The economics of electronic journals" in *Technology and Scholarly Communication* (eds Ekman, R. & Quandt, R.) 380–393 (Univ. Calif. Press, Berkeley, 1998). http://www.press.umich.edu/jep/04-01/odlyzko.html
2. Harnad, S. "Universal FTP archives for esoteric science and scholarship: a subversive proposal" in *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing* (eds Okerson, A. & O'Donnell, J.) 1 (Association of Research Libraries, Washington DC, 1995). http://www.arl.org/scomm/subversive/toc.html

The transition scenario

As soon as all refereed journal articles are self-archived by their authors in their institution's eprint archive, the literature is freed from all access barriers and impact barriers. Self-archiving could be done virtually overnight. The day after, all refereed research becomes freely accessible online to researchers the world over.

One possible outcome is that that will be the end of it. The refereed literature will be free online for those who want it and cannot get it any other way, but those who can afford to get it the old way via paying journals will continue to do so. In this event, the access/impact problem will be solved, but the library's budget crisis will not: it will simply become less important.

An alternative outcome is that when the refereed literature is accessible online for free, users will prefer the free version (as so many physicists already do). Journal revenues will then shrink and institutional savings grow, until journals eventually have to scale down to providing only the essentials (the quality-control service), with the rest (paper version, online PDF version, other 'added values') sold as options.

In none of these outcomes is peer-review itself compromised or put at risk; nor do authors have to give up, even temporarily, submitting to their established journals of choice. All they have to do is self-archive their preprints and postprints in their institutional eprint archives.

Nor are copyright restrictions an obstacle to self-archiving: preprints can be self-archived without any restriction at the time the paper is submitted to a journal. When the final draft is accepted, authors can ask the journal to retain their right to give away that draft online by self-archiving it. In practice, many publishers will agree to this if the author asks, although most do not publicly state it as policy. For these papers, the author can self-archive the refereed postprint alongside the pre-refereeing preprint(s). For those publishers who insist that all rights are transferred, authors can sign the agreement and self-archive a linked 'corrigenda' file, listing for the user what changes have to be made in the preprint to make it equivalent to the postprint. For details, see http://www.cogsci. soton.ac.uk/~harnad/Tp/resolution.htm#Harnad/Oppenheim.

© 2001 Macmillan Magazines Ltd

Setting logical priorities

A boycott is not the best route to free exchange of scientific information.

Ira Mellman

I am neither a publisher nor a professional editor. I am a practising scientist who, with 60 colleagues, collaborates with the nonprofit Rockefeller University Press to produce a high-quality publication, *The Journal of Cell Biology* (*JCB*; http://www.jcb.org). We do not aim to make a profit and have no vested interest in publishing. We do feel that, by helping to maintain a pre-eminent public forum for cell biologists worldwide, we are making an important contribution to our field, and to science in general.

I have long been an enthusiastic supporter of removing the barriers to free exchange of scientific information. The Public Library of Science (PLS; http://www.publiclibraryofscience.org) boycott initiative has the potential to provoke some important changes in this regard. As I and so many of my colleagues support the idea that journals should release their content six months following publication, the *JCB* has enthusiastically adopted this policy.

One would think that the biggest priority should now be to persuade other organizations, including the Nature Publishing Group (NPG; publisher of the *Nature* journals) and Elsevier Science (publisher of *Cell* and many other journals), to follow suit. Such commercial publishers have so far seemed ill-inclined to make their content free at any time after publication. This should be cause for considerable concern as together NPG and Elsevier produce some of our most widely read journals.

The PLS group has chosen a different and diversionary path. Not content to focus on ensuring public release of journal content, it has also demanded that released content be available for posting on any web server, anywhere. This demand appears to be poorly thought out, unnecessary, a waste of money and a potentially dangerous threat to scientific exchange forums such as the *JCB*. Moreover, to make such a demand by threatening a boycott of even those journals run by scientists for scientists strikes me as being needlessly anticollegial and counterproductive. This is not how science typically makes progress.

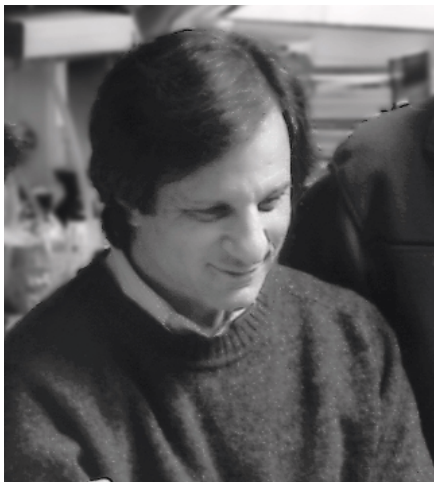
Why is the PLS demand for multiple-server release a bad idea? The nature of scientific information makes its reproduction delicate. It is not just plain text and sequences. Complications arise, for example in producing the myriad special symbols involved and in displaying complex visual images integral to many papers, particularly in cell biology. Each journal has its own translation software to take material formatted for publication of a print version and convert that material into a digital, online version. Frequent manual interventions are needed to correct the inevitable errors. Posting this information on each new site is fraught with the same difficulties. There is no software package or standard to ensure the accuracy of each re-posting, and the PLS has not thought through who will assume all the added costs.

Maintaining the integrity of the publication process is vital, in print or online. It is the responsibility of a journal to ensure that the science it publishes is a faithful, accurate and accessible version of the manuscripts it accepts for publication. It is not a process that should be ceded to unknown individuals or, in the case of PubMed Central (http://www.pubmedcentral.nih.gov), an agency of the US government.

Copyright conundrum

Viewed in this way, the issue of who owns copyrights is an ideological red herring, at least in the case of nonprofit publishers. As individual authors, we cannot protect our work from being re-posted on a multitude of servers, each with its own set of errors and deficiencies such as the loss of corrections, supplementary material and hyperlinks. Nor could we effectively monitor the repackaging of our work for resale by others. The Rockefeller University Press retains copyright of material published in the *JCB* not to make a profit, but to protect the integrity of the papers it publishes. Crucially, authors of *JCB* papers are free to reproduce their work for any scholarly or educational activity.

Many technical problems could be solved by investing money in software development. But is this necessary or advisable? Given that a revolution in scientific publishing was well under way even before PLS, what is the justification for the expenditure of public funds to duplicate efforts (such as the nonprofit HighWire Press) that are already successful? The end result would be a database lacking material from the preceding six months — of interest to few scientists. It would be far more logical to invest in increasing the power of websites such as PubMed (http://www.ncbi.nlm.nih.gov/PubMed) that permit searches on other sites without that material needing to be re-displayed. It is outmoded and incorrect to contend, as does the PLS, that hosting content on single sites is necessary for complete searching.



Extra effort: Ira Mellman says the burden of peer-review should not be underestimated.

It is unfortunate that the PLS boycott lumps together all types of journals: commercial and nonprofit, high-quality/high-circulation and archival or marginal-quality/low-circulation. It is perhaps a worthy goal to eliminate the last of these groups, many of which are probably poorly peer-reviewed, not widely read, and exist primarily to increase the profit margins of large-scale publishers. Their disappearance would save millions of dollars for libraries that are forced into subscribing to entire portfolios of such titles.

Ironically, the PLS demand for server release may pose the greatest risk to journals without large corporate sponsors, without government sponsors (as enjoyed by the *Proceedings of the National Academy of Sciences*) or without professional societies willing to bear chronic losses (for example, *Molecular Biology of the Cell*). If such journals, for example the *JCB* or the *EMBO Journal*, are forced to abandon their websites, their revenue and identities will be compromised, leading to the disappearance of the high-quality, well-read journals on which we rely for the best papers in our fields.

Even nonprofit, noncommercial journals need money to survive. After printing costs, the main expense faced by journals such as the *JCB* is the maintenance of an efficient reviewing system. When hard-copy journals finally disappear, our collective costs will drop; but high-quality and timely reviewing will still require a source of continuing revenue. Although imperfect, peer review has served the cause of science well. It can only increase in importance as the volume of scientific information increases. Guiding a scholarly and fair peer-review process is a challenge whose importance and complexity is underestimated by PLS as well as several other commercial online initiatives.

If readers were ultimately steered away from journal sites, journals would have insufficient funds to continue. If this risk were necessary to ensure the free exchange of scientific information, I would be advocating that we take it. The PLS initiative does not, however, make a persuasive case that the risk is necessary to achieve the desired result. Therefore, taking such a risk is not a logical approach to the problem. ■

Ira Mellman is chairman of the Department of Cell Biology, Ludwig Institute for Cancer Research, Yale University School of Medicine, PO Box 208002, New Haven, Connecticut 06520, USA.

© 2001 Macmillan Magazines Ltd

An instant in time

How the atomic clock revolutionized time-keeping.

Splitting the Second: The Story of Atomic Time
by Tony Jones
Institute of Physics: 2000. 199 pp. £14.99, $19.99

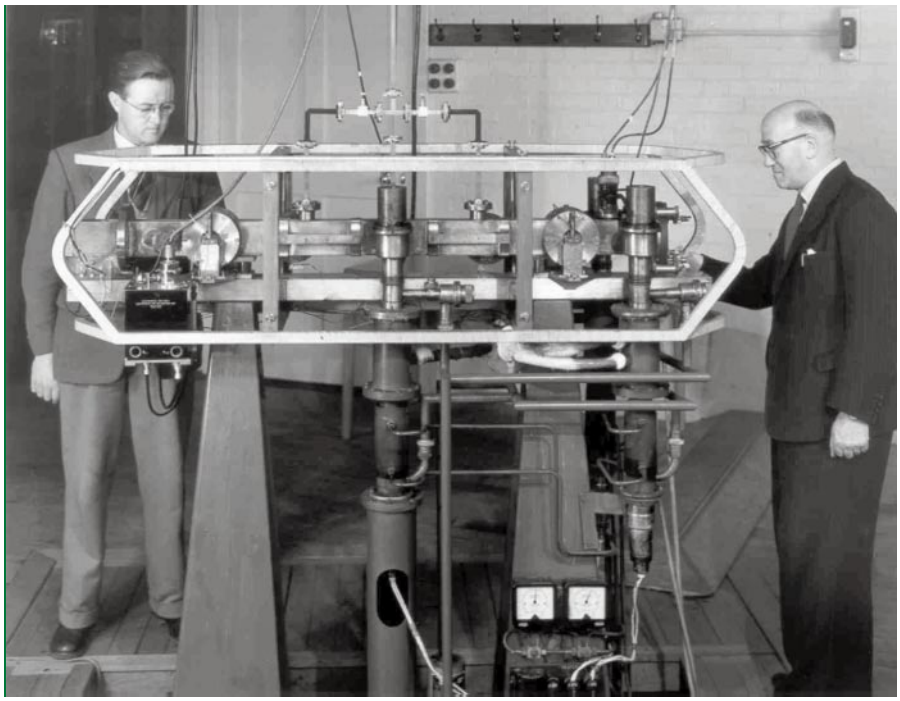
Daniel Kleppner

By the mid-1950s astronomical timekeeping had achieved unprecedented accuracy, based on a deepened understanding of planetary and terrestrial dynamics and the development of the quartz crystal clock. In 1952 the International Astronomical Union introduced a new timescale — ephemeris time, based on the motions of the planets — and in 1956 it redefined the second in terms of the length of a solar year. But the 1950s also saw the start of a revolution in timekeeping that in a single decade wrested the second from the astronomers and made ephemeris time obsolete.

The revolution was precipitated by the creation of the atomic clock, a device so unclock-like in appearance and operation that it was every bit as unpalatable to astronomers in the 1950s as John Harrison's mechanical chronometer was in the eighteenth century. *Splitting the Second* recounts the events of this revolution, the subsequent progress in atomic clocks and the many applications of timekeeping, few of which were dreamed of when atomic clocks were created.

The possibility of an atomic clock was first suggested in 1945 by the Nobel prizewinning physicist Isidor I. Rabi in a throwaway remark during a speech to the American Physical Society. A number of US groups picked up on his suggestion, but the first working atomic clock was created in the United Kingdom. In 1955, Louis Essen and John Parry operated a caesium atomic beam clock at the National Physical Laboratory in Teddington. The device was actually an atomic frequency standard, but the term atomic clock has become generally accepted. The Essen–Parry clock achieved an accuracy of one part in 10 billion (1 in 10^{10}). Refinements over the decades have increased the accuracy by a factor of close to 10,000. Recently, the fountain clock, a new type of clock that exploits laser-cooled atoms, achieved an accuracy of close to 1 in 10^{15}.

An obvious question is why anyone would want such high accuracy. The book addresses this question by demonstrating that as timekeeping improved, new applications steadily arose. The most spectacular of these is the Global Positioning System, which is finding fresh applications almost daily, from aiding the navigation of ambulances and cars to measuring the drift of continents.



NPL

Clocking on: Parry (left) and Essen with their caesium atomic frequency standard.

Accurate time is useless unless it can be disseminated, and *Splitting the Second* describes time dissemination in some detail. The echoes of history can still be heard, for instance, in the continued use of the term Greenwich Mean Time long after it became technically obsolete. The conflict between physicists' and astronomers' time can be detected in the international agreement for time dissemination. The time interval that is broadcast — the second — is physicists' time, controlled by an atomic frequency standard. But to satisfy the needs of astronomy, the timescale is adjusted to keep step with the slightly irregular motion of the Earth by introducing a leap second whenever the astronomical and physical times diverge by more than 0.9 seconds.

In one respect — reliability — atomic clocks cannot compete with astronomical timekeeping. To physicists, who generally care about measuring a frequency accurately rather than knowing the time itself, this is not of deep concern. But if one particular atomic clock, presumably the best in the world, were selected as the primary standard, sooner or later it would need to stop for repairs and the timescale would be broken. Consequently, primary atomic time standards are maintained not at one but at seven laboratories scattered around the world. These constantly compare their clocks both among themselves and with more than 200 secondary standards. The system is cumbersome but the timescale is accurate and reliable.

Jones recounts the historical events that underlie contemporary timekeeping and provides vignettes of the human factors involved in scientific and technological decisions. He makes a valiant effort to explain the science. His explanation of an atomic clock conveys the flavour of the physics and is probably as good as one can do in describing quantum phenomena in everyday language. Some details, such as the operation of a maser, are slightly confused, and some choices seem idiosyncratic. The name of Alfred Kastler, for instance, whose invention of optical pumping has played a continuing role in the development of atomic clocks, appears only in a list of Nobel prizewinners.

The book concludes with speculations on the future of timekeeping. These are not far-fetched, for the field is advancing rapidly. For instance, a method for measuring optical frequency, published too late to be described in the book, has opened the way to clocks that may one day achieve an accuracy in the range of 10^{18}. Anyone who is intrigued by fantastic precision or curious about how we really know the time, or simply likes tales of science and technology, will enjoy *Splitting the Second*. ■

Daniel Kleppner is in the Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.

© 2001 Macmillan Magazines Ltd

When space and time conspire

Complex Systems: Chaos and Beyond
by Kunihiko Kaneko & Ichiro Tsuda
Springer: 2001. 273 pp. $72

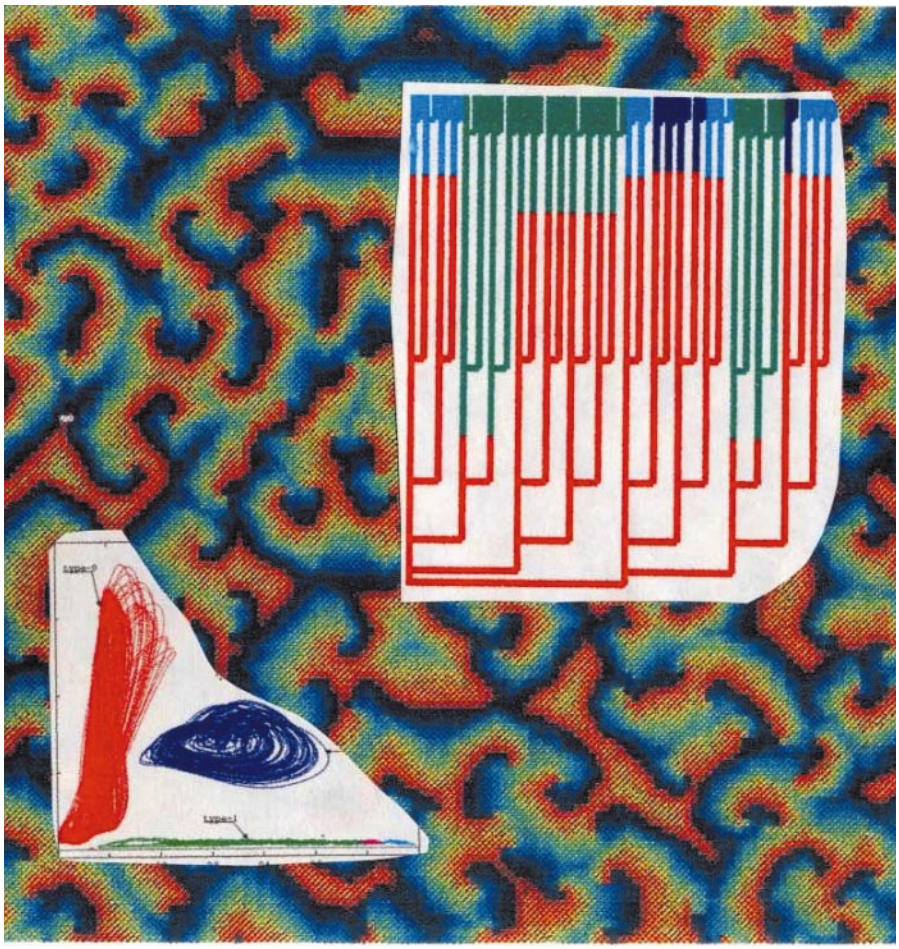
Michael F. Shlesinger

Books about chaos have the taste and flavour of regional cooking. Instead of Italian, Chinese, Lebanese or French dishes, one finds volumes devoted to different major themes of nonlinear dynamics. Some specialities include bifurcations, dissipative chaos, hamiltonian chaos and strange kinetics. Kunihiko Kaneko and Ichiro Tsuda bring to this mix their own brand of delicacy — spatiotemporal chaos. Their book focuses mostly on their own line of research, which is computer investigations of coupled map lattices, about which more later. Although equations are kept to a moderate number, the discussions are presented at a fairly advanced level. With careful reading, a graduate student might find a thesis topic or two.

Consider weather prediction. If it is cold and raining outside one's window, this provides a single point on a weather map. One could also keep a time series of the temperature. Although this might be informative, the single fixed observer would probably be unaware of irregular storm fronts and jet streams. In the same vein, the study of spatiotemporal chaos attempts to investigate and analyse spatially extended systems rather than just a single time series derived from measurements at one point. The quest is then to find what new phenomena can arise in spatiotemporal chaos that would remain hidden if one studied only temporal chaos. This is the core topic explored by Kaneko and Tsuda, together with suggested applications to biological networks and information processing in the brain.

One way of investigating spatiotemporal phenomena is by using a lattice of cellular automata that allow parallel computation. These automata are a lattice of cells that can take on numerical values, the simplest case being just zeros and ones. Each cell has an initial value and an update rule that determines for each time step the cell's next value. The rule is an equation that depends on a cell's present value and those of chosen neighbours. Although this appears simplistic, it can be shown that cellular automata can represent a universal Turing machine, meaning that it can theoretically calculate any quantity that is computable. In practice, one may not know how to set up and find the correct rules that would allow a cellular automaton to provide practical solutions to general physical problems.

Kaneko and Tsuda approach spatiotemporal chaos in the spirit of cellular automata. Rather than working with binary output, the value of a cell is calculated by iterating a nonlinear dynamical map that involves the previous value of a cell and a set of values of neighbouring cells. The numerical values produced by the map are capable of showing complex behaviour, including bifurcations and chaos in certain parameter ranges. Each map receives and sends information from and to neighbouring cells. This is the coupled map lattice model. Globally coupled maps have each cell receive a global input.



KUNIHIKO KANEKO

Order out of chaos: attractor portraits and cell lineage diagrams over a coupled map lattice.

These coupled map models can show a wide variety of spatial behaviours. With asymmetric one-way coupling between neighbouring cells the authors find spatial period doubling: that is, they find a set of periodic oscillations in values for upstream cells, followed by period-two oscillations and then period-four oscillations, and so on, as one moves spatially downstream. In coupled map lattices, as parameters are varied, spatial regions involving a subset of the cells can become frequency-, amplitude- or phase-locked. As these regions arise, they can stabilize, mix or disappear. As coupling strength is varied, parameter regimes can be found that produce regions that are spatially and temporally periodic, spatially quasiperiodic but temporally periodic, spatially chaotic but temporally periodic, and spatially and temporally nonperiodic. A statistical description of the above cases would be stable over time. It is also possible to find regions of spatiotemporal intermittency where the statistical description would change over time as distinct behaviours come and go. It is important to realize that all of these cases demonstrate emergent behaviours that are not programmed into the equations. Other possible dynamical behaviours include: propagating soliton turbulence where numerous correlated small spatial regions move with a constant velocity and collide and scatter with each other; noise-induced oscillator ordering where the addition of global noise provides a bandwidth of frequencies that facilitates the locking of oscillations among cells; and chaotic itinerancy where a trajectory intermittently changes its basic behaviour as it alternately falls under the influence of different unstable attractors. This dazzling array of behaviours hints at how complex the dynamics can be in spatially extended systems, as opposed to the temporal dynamical behaviour of a single nonlinear oscillator.

The coupled map lattice approach is not a detailed realistic model for any particular physical system, although the authors make intriguing speculative comparisons between generic dynamical behaviours and phenom-

© 2001 Macmillan Magazines Ltd

ena seen in real biological networks. Kaneko and Tsuda also seek ways of measuring the complexity of their systems. They define an information flow and entropy derived from the number and size of synchronized spatial clusters. And they make analogies between information storage, flow and processing in the brain and the many possible states of the coupled map lattice and the dynamical correlations among these states. Spatiotemporal dynamics of nonlinear coupled map lattices is a new field of study, and much remains to be done to determine whether it will be a fruitful approach to biological problems.

This is not a book for novices, but experts looking for stimulating discussions on spatiotemporal chaos and an in-depth tour of coupled map lattices will find it profitable. ■

Michael F. Shlesinger is in the Physical Sciences Division, Office of Naval Research, 800 North Quincy Street, Arlington, Virginia 22217-5660, USA.

Inspiration from new technology: *Macchia Forest* (1994) by Dale Chihuly.

Thinking the unthinkable

The Origins of Creativity
edited by Karl H. Pfenninger & Valerie R. Shubik
Oxford University Press: 2001. 268 pp. £20

Richard Gregory

Although we recognize creativity in the arts and sciences it is extraordinarily hard to define or explain. Why are some people — Mozart, Beethoven, Newton, Einstein — far more creative that almost everyone else? Is it a matter of genes — perhaps lucky combinations of inherited characteristics — or does it flower from a nurturing or challenging environment? What brain processes create creativity? These and many more questions are discussed here, by a dozen experts, with a useful introduction and analysis by the editors. The contributions vary from personal reminiscences to summaries of a life's work. The very wide variety of approaches — experiences in art and science, the biological basis and brain mechanisms of creativity, emotion and reason, the role of the environment, perception of patterns — reflects the lack of an accepted paradigm for considering creativity.

Thomas Cech starts the discussion with charm and generosity, by relating his Nobel-prizewinning discoveries on self-splicing RNA and its implications for the origin of life, to the combined creativity of his laboratory team. This emphasizes the importance of secure long-term support for creative science, to nurture shared confidence and loyalty. These are not bought cheaply, but yield results.

Dale Chihuly, a remarkable artist who works with glass, also stresses the importance of teamwork, together with the inspiration and possibilities of new technologies. Physician and wood-carver David Rogers sees leading a group as a creative activity, even though his female-form art is private. Relations between art and science are discussed by Gunther Stent, pioneer in molecular biology and philosopher of science, who points out interesting parallels, and some non-parallels, between discovering and creating.

The distinguished neuropsychologist Antonio Damasio reverses the long philosophical and scientific tradition of separating reason from emotion by suggesting that emotion is essential for creative rational thinking. He presents a detailed account of brain anatomy and cognitive processes at the leading edge of brain research. The emphasis is on novel and tentative representations in working memory, for trying out possibilities and new combinations, driven by emotional pushes and pulls. This is eloquent and compelling.

The composer Bruce Adolfe discusses musical imagination related to brain activity and working memory, with advice on how to be creative. Karl Pfenninger looks at how the nervous system might generate the 'higher function' of creativity, stressing the brain's plasticity and the importance of early environment and learning. The importance of early experience is developed by Janina Galler, using the interesting results from lengthy experiments on the effects of malnutrition on intelligence and creativity.

Howard Gardner, the well-known expert on the psychology of intelligence, discusses his 'multiple intelligences' experiments and ideas, using the example of seven exceptional people who have affected other people's lives. The 'cases' are few but the analysis is deep, with detailed biographical scholarship. Gardner relates effects of society to brain function and to different kinds of intelligence. The cell biologist George Palade discusses "the interactions between the arts and sciences in the context of Western civilization as a function of time in history", stressing the advantage of affluence in the society and a Protestant work ethic in the home background.

Introducing the subject of patterns of perception, painting is discussed by the artist Françoise Gilot, who describes her relationship with Picasso and her own experience with paint and canvas. The theme of vision is steered by the neurobiologist Charles Stevens into how eyes and brains work. Structures and mechanisms of the retina and visual cortex are important, but perhaps limitations of space prevented him voyaging into cognitive (knowledge-based) processes of vision, which seem more directly relevant. Indeed, is human creativity just a simple extension of the creativity of all perception — as even everyday perceptions are so much richer than the available information?

This leads the reader to the essay from the great Benoit Mandelbrot, and fractals. They evoke feelings of creative art — but are they works of art? Where is the artist in a computer algorithm? This intriguing discussion might have led to the subject of creativity — or the lack of it — in computers, but, perhaps sadly, it does not.

The book ends with an editorial synthesis and critique of the contributions. Perhaps this will make the reader question the notion of creativity. Is creativity so specifically human, as implied here? Will computers beat us at science, even at art? Surely, darwinian evolution, though lacking in emotion, is the most creative process we know? Shouldn't this rather well-understood process be discussed here? Admittedly, natural selection takes its time, but just look at the results. ■

Richard Gregory is in the Department of Experimental Psychology, University of Bristol, Bristol BS8 1TN, UK.

© 2001 Macmillan Magazines Ltd

Measuring the past

All Done with Mirrors
by J. F. Neal
Secret Academy Press: 2000. 274 pp. £25, $38

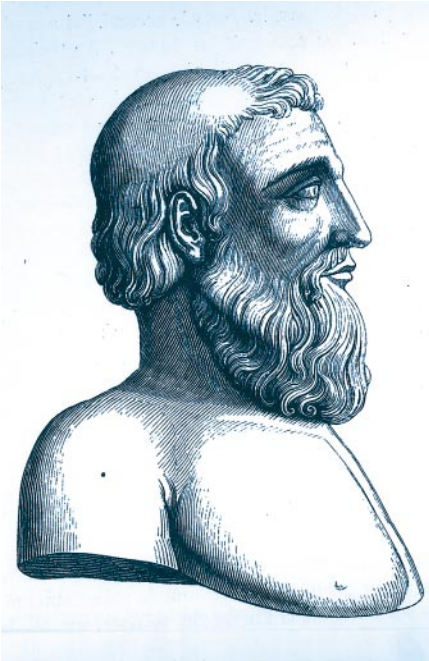
Michael Vickers

Thanks to John Neal's remarkable new book, ancient metrology — once the playground of Newton, but now largely ignored even by archaeologists — should cease to be a pariah subject and regain its place at the centre of the study of antiquity. In the past, the widely attested variations in ancient reported linear measurements have been put down to sloppiness on the part of our ancestors. But Neal is able to show that such variations belong to a logical, elegant and cohesive system partially based on divisions of the Earth's surface at different points on the longitudinal meridian.

The Earth is not a true sphere, but is subject to polar flattening, which means that the longitudinal meridian, or the great circle through the poles, is essentially elliptical. The distortion is minute, but it creates a measurable variation from degree to degree. Degrees nearer the poles are longer than those at the Equator. Thus, a widely accepted value for the Greek foot of 1.0114612 feet proves to be 1/360,000th part of the longitudinal meridian degree at just under 38° latitude, the same latitude as that of the Aegean.

Elaborating a scheme first noted by the philosopher and historian John Michell, Neal observes that feet (or cubits) stand in a ratio of 175:176 to larger units in a series. This at once explains the Roman architect Vitruvius' account of an odometer — an instrument for measuring the distance travelled by a wheeled vehicle — that contained a mechanism designed audibly to release a stone into a box every mile, in his case, 400 revolutions of the 12½-foot-perimeter wheel to the 5,000-foot mile. If Vitruvius' 4-foot radius to 12½-foot perimeter, or 3.125π ratio, was strictly adhered to, there would have been a discrepancy of more than 28 feet in every mile. But if the shorter Roman foot of 0.967680 feet was used for the diameter of the carriage wheel, and the longer Roman foot of 0.973209 feet used for the perimeter, the calculation of the mile is accurate in terms of the longer measure. The difference between $^{22}/_{7}$ and $^{25}/_{8}$ can be expressed as 3.142857 = 176, and 3.125 = 175 (both values of π were used in the ancient world).

Neal notes that if a diameter is a multiple of either four or eight, 3.125 may be accurately used to maintain an integral number in the perimeter, because the ratio between using true π as the module of measurement of the diameter is the 175th part less than that of the perimeter. There was thus a practical



Marcus Vitruvius Pollio: not missing the mile.

MARY EVANS PICTURE LIBRARY

purpose underlying variational fractions between the ancient standards (and this is but one of many), and they can no longer be put down to carelessness or error.

Although the system is complex, it is blindingly obvious once it is tabulated. Roman, Greek, Egyptian and Babylonian measures are seen to be interrelated. Life is breathed again into the late Alexander Thom's 'megalithic yard', a unit of measurement Thom found to have been consistently used at many prehistoric megalithic sites. As Neal points out, "not only is the megalithic system largely ignored by archaeologists, it is opposed — even by the numerate among their ranks". This position is now untenable, as it can be shown that the megalithic yard shared an origin with the Sumerian cubit. And the foot-measure used in England — equivalent to a Greek foot — proves to have played a pivotal role in the whole metrological system. It is ironic that just as it is being thrown on the scrap heap of history, its historical importance is beginning to be recognized.

Another problem besetting metrology is that it has become the lodestar of internerds and pyramidiots. Neal is extremely careful of the feelings of mathophobes, and develops his ideas in a way that the innumerate can follow and the mathematician enjoy; he gives short shrift to "pyramid-inchers" and "inch-prophecy theories" (look up, but not for long, the Great Pyramid on the web). By contrast, this is a sober and thoughtful analysis with far-reaching consequences for the study of the past. More than that, it is a major contribution to the history of science. ■

Michael Vickers is in the Department of Antiquities, Ashmolean Museum, University of Oxford, Oxford OX1 2PH, UK.

A passion for butterflies

A nineteenth-century collection of the Large Blue (*Maculinea arion*), which became extinct in Britain in the late 1970s, one of the many fine illustrations of butterflies and those who studied them in *The Aurelian Legacy: British Butterflies and their Collectors* (by Michael Salmon, Harley Books: 2000, £30.00). The story of the Large Blue well illustrates the mania that gripped some nineteenth-century collectors, with the taking of "more than 2,660 specimens" of this rare butterfly at a Cornish site in 1896. Most of the entomologist-collectors described by Salmon were fortunately less acquisitive and laid the foundations of our present knowledge of butterfly biology and ecology. This information-packed but highly readable account of 300 years of British lepidoptery, practised largely by amateurs, is a must-read for butterfly afficionados and social historians alike.

The amateur tradition continues, and records collected by thousands of volunteers between 1995 and 1999 provided the data for *The Millennium Atlas of Butterflies in Britain and Ireland* (edited by Jim Asher *et al.*, Oxford University Press: 2001. £30.00, $40.00). This handsomely produced book records the most comprehensive survey of butterflies ever undertaken in Britain and Ireland, and provides an invaluable picture of the state of all the native species, now under threat from habitat destruction if no longer from the killing bottle and collecting box. Happily, the Large Blue figures in this book as well, having been reintroduced to managed sites in the 1980s and 1990s.

© 2001 Macmillan Magazines Ltd

Scientist's birthright

How a new name embodied ideals of connection and inclusiveness.

Dennis Danielson

Although histories of science routinely mention the birth of our word 'scientist' from the pen of William Whewell a mere 167 years ago, revisiting the social and literary moment of that birth offers a fresh glimpse of the word's worthy parentage and generous potential.

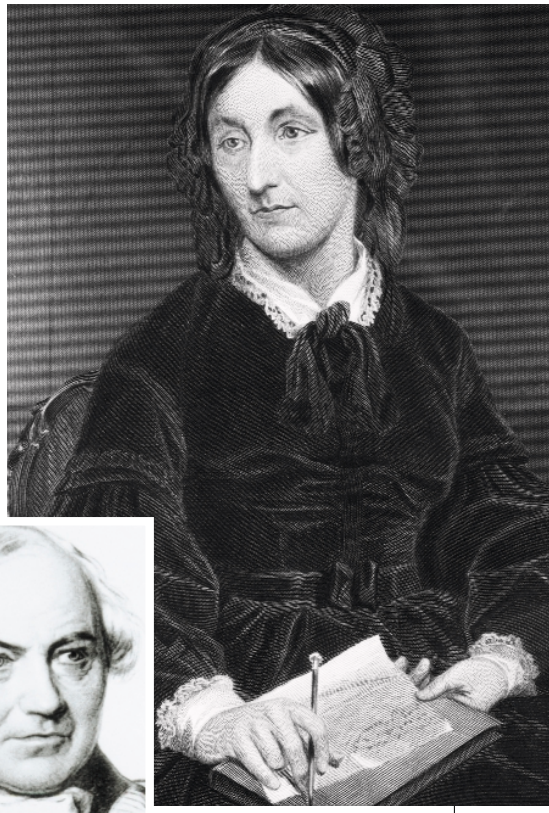
The social events most directly related to Whewell's invention of 'scientist' were the first meetings of the British Association for the Advancement of Science (BAAS), in York (1831), Oxford (1832) and Cambridge (1833). The earliest of the BAAS proceedings reveal the strikingly inclusivist aims of the new association. In contrast with the conspicuously exclusive, London-based Royal Society, the BAAS (in the welcoming words of W. V. Harcourt) gathered "many distinguished members of learned and scientific bodies" from around the United Kingdom "in consequence of a general invitation to the friends of science".

It was, significantly, from a provincial body (the Yorkshire Philosophical Society) that the invitation issued. And against the recognition that "in our insular and insulated country, we have few opportunities of communicating with the cultivators of science in other parts of the world", the purpose of the BAAS was to open "new channels of communication" both internally and internationally, as well as to "promote science in every part of the empire." Moreover, it was decided that, at the meetings themselves, the "least abstract" of the scientific papers should be presented in evening sessions, to which would be admitted "a more popular audience".

All this emphasis on widening connections needed to be matched by a breaching of disciplinary boundaries. Warning that specialization can lead to "insulation", Harcourt declared: "The chief Interpreters of nature have always been those who grasped the widest field of inquiry, who have listened with the most universal curiosity to all information, and felt an interest in every question which the one great system of nature presents. Nothing ... could be a more disastrous event for the sciences, than that one of them should be in any manner dissociated from another."

And yet, of course, nowhere does there appear a single name for the citizens of the expanding commonwealth of science that Harcourt envisages. There are "friends of science", "cultivators of science", and "interpreters of nature" — but still no scientists.

The literary context in which that word first appears is Whewell's anonymous essay (in *The Quarterly Review*, 1834) on Mary Fairfax Somerville's book *On the Connexion of the Physical Sciences*. This context is highly germane to the meaning of the new word, as signalled by Somerville's title with its stress on "connexion" — a motif which Whewell eagerly marshals against the forces of disintegration. This danger he renders in metaphors of bodily, territorial and imperial fragmentation, lamenting "the tendency of science ... to separation and dismemberment"; its "disintegration ... like that of a great empire falling to pieces"; the "division of the soil of science into infinitely small allotments". He acknowledges the contribution of the BAAS in "bringing together the cultivators of different departments", and he warmly praises Somerville for attempting in her own work to achieve the same end.



Connected: inspired by Mary Somerville, William Whewell coined the word 'scientist'.

SPL/CORBIS

It is in the midst of this discussion of the need to connect the sciences that Whewell refers to those first meetings of the BAAS, which, he says, "felt very oppressively the want of any name by which we can designate the students of the knowledge of the natural world collectively". He recounts how terms from other languages were rejected: "*Savans* was rather assuming, besides being French instead of English" and the German *Natur-forscher* "might suggest such undignified compounds as *nature-poker*, or *nature-peeper*"! But sandwiched between these foreign nonstarters is Whewell's suggestion — made by a person he identifies merely as "some ingenious gentleman" — "that, by analogy with artist, they might form scientist". However, he adds (a little prematurely, as things turned out), "this was not generally palatable".

Clearly, the literary and social moment in which 'scientist' sprang forth was pregnant with a longing for greater "connexion" across various boundaries, including those of territory and discipline. It was a motive that continued to shape many scientific efforts in the decades following 1834 — for example, those of another admirer of Somerville's: Alexander von Humboldt, author of *Kosmos* and founder of what came to be called ecology, a science whose very foundation is the interconnectedness of disciplines and of living things and their environment.

> The chief interpreters of nature are those who have grasped the widest field of inquiry.

However, it is worth noticing that Somerville's quickening presence at that vital moment marks two other crucial dimensions of inclusiveness that the name 'scientist' has not always consistently embodied. First, as Whewell implies, Somerville modelled with great lustre what we now, still rather nervously, call popular science: "How valuable a boon it is to the mass of readers, when persons of real science, like Mrs Somerville, condescend to write for the wider public."

The other boundary to be breached — which Somerville eminently did breach — is that of exclusion by gender. As Whewell says, with fitting satiric tone, "there are few individuals of that gender which plumes itself upon the exclusive possession of exact science, who may not learn much" from this book by a woman.

Such, at least in promise, was the inclusive birthright of 'scientist'. Long — and increasingly — may that birthright be exercised. ■

Dennis Danielson is editor of The Book of the Cosmos *(Perseus Publishing, 2000) and professor of English at the University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada.*

© 2001 Macmillan Magazines Ltd

The genomic cosmos

Mark C. Fishman

The genetic screen provides an ideal tool for establishing a pan-genomic phenotypic compendium. Experimentally, it begins by showering germ cells with a mutagen such as a chemical, radiation or virus. Over ensuing generations, appropriate breeding schemes bring the mutant phenotype into view, and those of interest are selected and isolated. All regions of the genome are believed to be equally accessible to many mutagens. Smaller genes may be hit less often, but ultimately none is invincible if the screen is large enough. Genetic screens, if of sufficient size, therefore examine the roles of all genes in a genome.

It's amazing that the genetic screen works at all. First, at least in higher metazoans, most genes are used in multiple locales and several stages of development. A phenotype of interest may be buried and confounded by widespread disarray, or not even have a chance to appear because of premature death. Second, many important processes have molecular back-ups, so redundant systems compensate for individual mutations.

But screens have succeeded. First, when done in a large-scale way, to 'saturation', by their all-inclusiveness they have provided a logic, a sense of what can go wrong. Perhaps this is not surprising for single-cell organisms, but many were surprised by just how informative were large-scale screens done on the fruitfly *Drosophila*. These screens proved that the parts of the body plan are not so inextricably linked to each other as to prevent analysis of embryonic development. Individual mutations can perturb single embryonic segments, or patterns of segments, without causing total chaos. The large-scale genetic screens in zebrafish reveal modular elements even to the assembly of vertebrate organ form and function, such that a mutation can delete one chamber of the heart but leave the rest of development relatively untouched.

This biological logic is revealed by screens even before the cloning of the mutant genes. That embryonic body plan can be dissected as a hierarchy of axial and segmental pathways was intuited from mutant phenotypes isolated in a large-scale *Drosophila* genetic screen well before discovery of the molecular identity of the mutations.

The second level of success occurs after mutation cloning. Each mutation discovered can provide an entrance point to a biochemical pathway. The other components of the pathway may have eluded the screen because their mutations caused no informative effects. But having the one molecular handle provides the needed molecular entrance point. Other elements are accessible by other methodologies. The biochemical intricacies of cell death, for example, have been revealed over the years by biochemical or genetic means in many species, but originated in the genetic screen-based revelation of the *ced* genes in the nematode worm *Caenorhabditis elegans*.

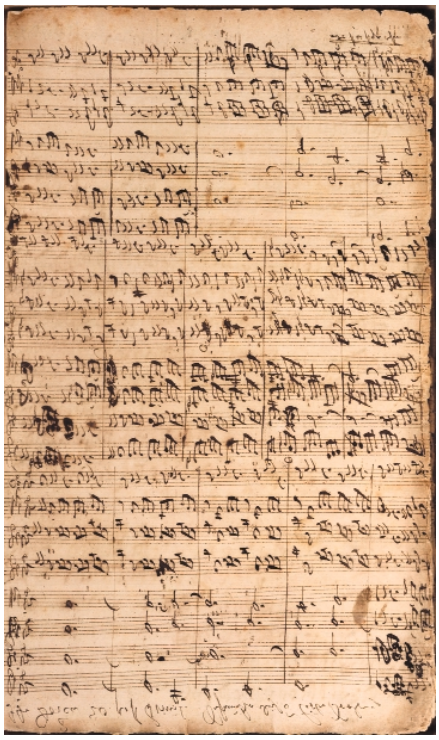
The interpretation, and indeed power, of a screen is observer-dependent. There are many possible thematic interpretations of a screen, all woven together as in an Escher print or Bach canon, and none is most 'correct', although some can provide more experimentally useful predictions than others. Even the defined phenotype for any given mutation reflects but one part of the function of the gene, the one that happens to be most prominent or fall in the field of focus of that particular scientist. Rashomon-like, what is deemed distinctive and important by one observer may go unnoticed by another.

The name given to the mutation reflects this bias of the discoverer. Provision of the gene's moniker may seem poor recompense for hours spent sorting and breeding embryos. But, as Umberto Eco noted of Adam's role in Genesis, there is tremendous power to the Nomothete, the name-giver, the person who creates the language to describe the biological universe. Unlike Adam, the screener's chosen name rarely reflects divine inspiration. Some, such as *notch*, may coldly reflect the phenotype; others, such as *groucho* or *heart of glass*, taste in comedians or rock and roll. Hence, embedded in the name is a bit of scientific folklore, seeming flippant to the formal-minded but endearing to the *cognoscenti*.

What is the future of screens? Methodologically, some will be less random. With all genes defined and, at least in some species, tools to incapacitate them one-by-one (RNAi, morpholino antisense), it will be possible to march along chromosomes percussing each gene for phenotypic resonances. Chemicals can phenocopy mutations and even reveal pathway components that are not accessible in genetic screens, and large-scale screens using small molecules are under way. In terms of the biological targets of screens, so far most have been related to embryonic development. But there are many opportunities for expansion into other arenas. There are already hints from *Drosophila* and zebrafish that the ontogeny of physiological functions, heart-rate control for example, may be tractable to screen methodology. Obviously, molecular understanding of essential function and homeostasis also generates candidate genes for complex diseases. In so far as physiological adaptability provides an umbrella for mutation accumulation, these screens also may illuminate essential components of evolution.

Thus, like Borges seeking to classify and order his universe, which he likened to a library of Babel, our goal no longer is to catalogue the letters on the spines of the books, which we know from the genome projects, but rather to decipher the volumes' phenotypic content. Recent successes indicate that future screens will indeed reveal the overarching logic, order and component units to this universe of phenotypic possibilities. ■

Mark C. Fishman is in the Harvard Medical School and Massachusetts General Hospital, Fruit Street, Boston, Massachusetts 02114, USA.



BRIDGEMAN ART LIBRARY/CHRISTIE'S

In tune? Interpretations of genetic screens can be woven together as in a Bach canon.

Genetic screens

"An organism's genome is but a pale non-homuncular image of its universe of biological functions. How can we scan the true biological universe, that of the cosmos of phenotypes?"

FURTHER READING

Nüsslein-Volhand, C. & Wieschaus, E. *Nature* **287**, 795–801 (1980).

Fishman, M. C. & Olson, E. N. *Cell* **91**, 153–156 (1997).

Borges, J. L. *The Library of Babel* in *Labyrinths* (New Directions Publishing, New York, 1962).

Eco, U. *Serendipities: Language and Lunacy* (Harcourt, San Diego, 1998).

Ellis, H. & Horvitz, H. R. *Cell* **44**, 817–829 (1986).

© 2001 Macmillan Magazines Ltd

Polio vaccines exonerated

Robin A. Weiss

One theory about the origin of the AIDS pandemic is that the virus responsible, HIV, was transmitted to humans from chimpanzees through contaminated polio vaccine. That theory fails some crucial tests.

Vaccine manufacturers need to be ever vigilant lest their products become unwittingly contaminated by microbes lurking in the source materials used to make vaccines[1]. It was with such a possibility in mind that in 1999 Edward Hooper[2] promulgated a startling theory about how the human immunodeficiency virus type 1 (HIV-1) first came to infect humans. Hooper proposed that early batches of live oral polio vaccine (OPV), supposedly grown in chimpanzee kidney cell cultures at the Wistar Institute, Philadelphia, and tested in Africa, became contaminated with a chimpanzee virus that was able to infect its new host — with lethal consequences.

Three brief papers[3–5], beginning on page 1045 of this issue, bring in a 'not guilty' verdict on OPV. Tests by Blancou *et al.*[3] and Berry *et al.*[4] on early OPV stocks prepared at the Wistar Institute in the 1950s show that they were propagated in the kidney cells of rhesus monkeys, and that they lack nucleic-acid sequences related to either HIV or chimpanzees. And Rambaut and colleagues' evolutionary analysis[5] of HIV-1 subtypes from the Democratic Republic of Congo (Zaire) indicates that the common ancestor to HIV-1 group M, which gave rise to the pandemic strains, was present in a human host long before the first OPV field trials were conducted in the Congo during the late 1950s.

The idea that HIV might have been introduced to humans through contaminated polio vaccines has been around a long time. I first learned about it from an antivivisectionist tract circulated in 1987. The theory became notorious through Tom Curtis's article[6] in *Rolling Stone* in 1992, which proposed that HIV came from a simian immunodeficiency virus of African green monkeys (SIV_{AGM}). By 1963, production of OPV had switched from using cells from Asian macaques to those from African green monkeys, following the discovery of another virus — simian vacuolating virus 40 — as a contaminant in macaque tissue[1,7]. Twenty-five years later, many African green monkeys were found to harbour SIV_{AGM}, necessitating screening for this virus or use of human cell substrates in vaccine preparation.

But SIV_{AGM} is only distantly related to HIV-1, and cannot have been its precursor. The closest known animal virus to HIV-1 is SIV_{CPZ} of chimpanzees; in fact, the three main groups of HIV-1, named M, N and O, differ from each other as much as they differ from SIV_{CPZ}, suggesting that there were at least three separate instances of cross-species transmission[8]. So if the OPV hypothesis on the origin of HIV were to remain plausible, the use of chimpanzee kidneys for OPV propagation had to be invoked[2]. Yet there is no documentation that chimpanzees were ever used for this purpose and, along with a paper simultaneously published in *Science*[9], these new reports[3–5] dismiss what reasonable conjecture existed.

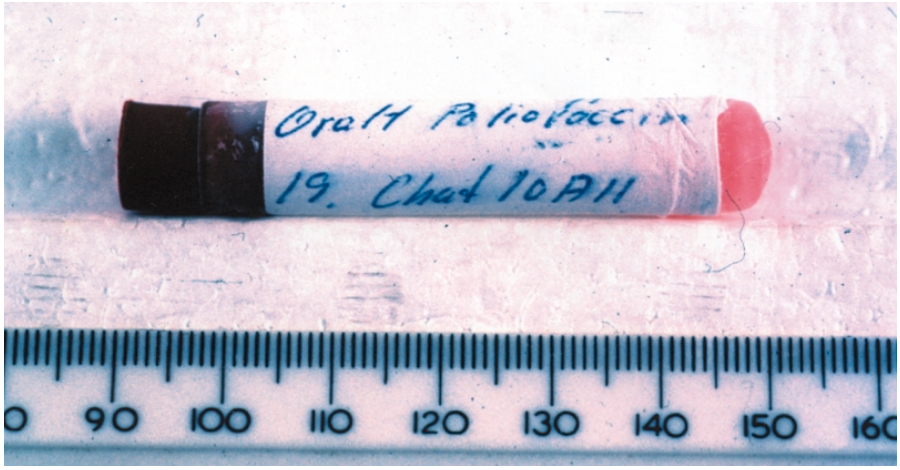


Figure 1 **An original vial of CHAT 10A-11 oral polio vaccine (OPV). The vial was found at the National Institute for Biological Standards and Control, Potters Bar, UK. New tests show that the vaccine was prepared from the tissue of rhesus macaques, not of chimpanzees as required by the OPV-transmission hypothesis. (Photo courtesy of Harvey Holmes[4].)**

From the sensitive and specific forensic DNA tests based on the polymerase chain reaction, it is clear that the cellular substrate for the early OPV batches was macaque tissue. This has been determined by three independent laboratories[3,9] in samples made available by the Wistar Institute, which included the OPV batch (designated CHAT 13) used in Léopoldville (Kinshasa) between 1957 and 1959.

The OPV batch that Hooper[2] considered to be under most suspicion, however, was CHAT 10A-11, which was also tested in Belgium's African colonies in the late 1950s. This batch was not available from the Wistar, but an original vial of the batch was found at Britain's National Institute for Biological Standards and Control (Fig. 1), and the new tests show that it was prepared from rhesus-macaque cells[4]. Further tests showed that these batches contained type-1 poliovirus only, as originally recorded for CHAT vaccines. There is no reason to doubt their authenticity.

Last year, Korber and colleagues[10] extrapolated the timing of the origin of HIV-1 group M back to a single viral ancestor in 1931, give or take about 12 years for 95% confidence limits. Because this calendar of events obviously pre-dated the OPV trials, in the revised version of his book[2] Hooper suggested that group M first began to diverge in chimpanzees, and that there were then several independent transfers of virus to humans via OPV. In that case, several OPV batches should bear evidence of their production in chimpanzee tissue, yet no such evidence has been found[3,4,9]. In any case, the data of Rambaut *et al.*[5], and those described in a forthcoming paper by Sharp *et al.*[11], indicate that Hooper's interpretation is vanishingly unlikely. Rambaut *et al.* re-analysed sequence data, reported last year[12], on the complexity of Congolese HIV-1 group M isolates. The diversity of HIV-1 in the Congo appears to pre-date the evolutionary radiation or 'starburst' of the major virus subtypes that have since spread across the world. But the genetic sequences conform to a single origin, presumably 'patient zero', the first human to become infected.

Defenders of the OPV hypothesis hold that the alternative route of cross-species virus transfer, an accidental infection from a chimpanzee, seems equally unlikely. Most

© 2001 Macmillan Magazines Ltd

of today's human infections, however, have come from various animal sources within the past 10,000 years or so[13], a split-second on the evolutionary timescale. So SIV_{CPZ} may have infected humans on numerous occasions in human history without becoming firmly established in our species, rather like the Lassa fever and Ebola viruses. What may have helped HIV-1 group M to take off epidemically could have been the use of non-sterile needles and syringes in Africa in the mid-twentieth century[14]. This mode of transmission is currently behind the explosive spread of HIV-1 in parts of Russia and China. So one can envisage the involvement of an unintended medical factor in the adaptation of HIV to humans after it had crossed species from chimpanzees.

The new data[3–5,9] may not convince the hardened conspiracy theorist who thinks that contamination of OPV by chimpanzee virus was subsequently and deliberately covered up. But those of us who were formerly willing to give some credence to the OPV hypothesis will now consider that the matter has been laid to rest. One may expect to hear an argument that chimpanzee kidney cells could have been used locally in Africa to amplify the batches of OPV prepared at the Wistar Institute in macaque cells, and that it was these vaccine samples that became contaminated. But the facilities for this type of cell culture did not exist in the Congo in 1957–59; besides, the evolutionary analysis of HIV-1 group M contradicts such a view.

The irony is that these new studies would almost certainly not have been undertaken if Hooper[2] had not called for the analysis of DNA in stored OPV stocks, on the suggestion of the late Bill Hamilton. So we owe Hooper and Hamilton a debt of gratitude for pressing the case for those tests. When the preliminary results of the investigations were announced at the Royal Society in London last September, Hooper nonetheless dismissed them as "irrelevant to the OPV hypothesis". But that simply isn't so — on the contrary, some beautiful facts have destroyed an ugly theory. ■

Robin A. Weiss is in the Department of Immunology and Molecular Pathology, University College London, 46 Cleveland Street, London W1T 4JF, UK.
e-mail: r.weiss@ucl.ac.uk

1. Weiss, R. A. *Emerg. Infect. Dis.* **7,** 153–154 (2001).
2. Hooper, E. *The River: A Journey Back to the Source of HIV and AIDS* (Penguin, London, 1999; revised edn, 2000).
3. Blancou, P. *et al. Nature* **410,** 1045–1046 (2001).
4. Berry, N. *et al. Nature* **410,** 1046–1047 (2001).
5. Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M. & Holmes, E. C. *Nature* **410,** 1047–1048 (2001).
6. Curtis, T. *Rolling Stone* 54–60, 106, 108 (19 March 1992).
7. Sweet, B. H. & Hilleman, M. R. *Proc. Soc. Exp. Biol. Med.* **105,** 420–427 (1960).
8. Gao, F. *et al. Nature* **397,** 436–441 (1999).
9. Poinar, H., Kuch, M. & Pääbo, S. *Science* **292,** 743–744 (2001).
10. Korber, B. *et al. Science* **288,** 1789–1796 (2000).
11. Sharp, P. *et al. Phil. Trans. R. Soc. Lond. B* (in the press).
12. Vidal, N. *et al. J. Virol.* **74,** 10498–10507 (2000).
13. Diamond, J. *Guns, Germs and Steel* (Vintage, London, 1998).
14. Marx, P. A., Alcabes, P. G. & Drucker, E. *Phil. Trans. R. Soc. Lond. B* (in the press).

Palaeontology

Ruffling feathers

Hans-Dieter Sues

The evolution of feathers and flight were generally thought to be inextricably linked. But new fossils from China show that feathers pre-dated the origin of flight and of birds.

One of the liveliest debates in palaeontology has concerned the origin of birds. There is now overwhelming anatomical evidence that birds evolved from small predatory dinosaurs known as theropods, and that the two groups are linked by a remarkable series of transitional forms[1]. Most palaeontologists accept this evidence. Only a small (if vocal) group continues to argue that birds have no close relationship to dinosaurs[2].

Feathers are the most distinctive attribute of living birds. Traditionally, their evolution has been linked to the origin of flight[2], but there have always been a few dissenting opinions[3,4]. Two new studies, one published last month[5] and the other in this issue[6], now confirm that true feathers already existed in the non-flying dinosaurian relatives of birds and thus pre-dated the origin of birds and avian flight.

In recent years, there have been several reports of feathers and feather-like integumentary structures in various non-avian theropod dinosaurs from the Yixian Formation of Liaoning province in northeastern China[7–9]. The Yixian Formation is of Early Cretaceous age (dated at 125 million years ago) and consists of a series of layered lake sediments and volcanic ashes. These deposits contain many remarkably preserved fossils of an amazing array of terrestrial, flying and aquatic animals. The fine-grained sedimentary rocks often retain traces of soft tissue and, in some cases, even identifiable gut contents.

One of the theropod dinosaurs from the Yixian Formation, *Sinosauropteryx*[7], showed unbranched fibre-like structures fringing the back of the head, neck, back and tail, which some researchers considered to be proto-feathers but others dismissed as frayed internal fibres of collagen, a structural protein found in connective tissue. However, two other theropod dinosaurs from the same deposits, *Caudipteryx* and *Protoarchaeopteryx*[8], undoubtedly possessed true feathers. In both, the body was covered by small feathers. *Caudipteryx* also had primary feathers attached to the second (longest) finger of its hands and sported a fan of feathers at the end of its tail. The tail of *Protoarchaeopteryx* bore symmetrical, vaned feathers in a fan-like arrangement. Opponents of the theropod–bird connection have suggested that *Caudipteryx* was actually a flightless bird[10], but this claim is not supported by its skeletal structure.

The critics also noted that no feathers were known in dromaeosaurs, which share scores of derived skeletal features with early birds and thus are generally considered to be their closest relatives[11]. Dromaeosaurs were a group of small- to medium-sized, non-avian theropods, which include *Velociraptor*, the reptilian villain in the blockbuster movie *Jurassic Park*. They were especially characterized by the presence of a greatly enlarged claw on the second toe of the foot. Dromaeosaurs had long arms and long, grasping hands, but their forelimbs were not modified into wings.

In a report published last month, Xu *et al.*[5] documented the presence of compound filamentous integumentary structures in the somewhat scattered but well preserved remains of the type specimen of the dromaeosaur *Sinornithosaurus millenii*[9] from the Yixian Formation. These structures exhibit two kinds of feather-like branching: filaments joined in a basal tuft, and several filaments joined at their bases in series along a central filament. What makes them particularly remarkable is their close correspondence to transitional stages II and IIIA predicted by a developmental model[12] for the evolutionary origin of feathers in birds (Fig. 1).

In the second of the new papers, Ji *et al.*[6] (page 1084 of this issue) report on the latest remarkable find from the Yixian Formation — a nearly complete, articulated skeleton of an unidentified dromaeosaur, most of the body of which is densely covered with feather-like structures. The complex filamentous structures on the arms and tail of the new fossil are especially striking. Most of them exhibit a radiating pattern of filaments originating from a single point. The structures on the arms appear to show a herring-bone pattern around a central stem, which is similar to bird feathers with their central rachis and serially attached barbs. Feathers are the only kind of covering found on living vertebrates that shows such a branching structure. So Ji *et al.*[6] explicitly equate the filamentous structures of the new dromaeosaur to the feathers of birds.

The exceptionally preserved fossils from

© 2001 Macmillan Magazines Ltd

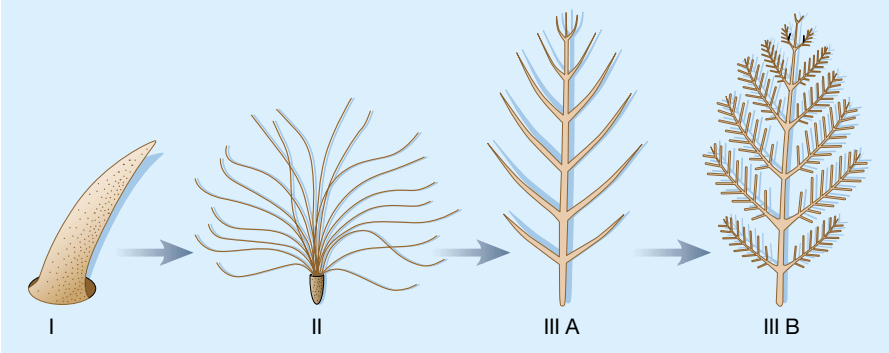


Figure 1 **Hypothesized stages I–III of feather evolution[12]. Stage I of this model assumes an unbranched, hollow filament, which developed from a cylindrical invagination of the epidermis around a papilla. In stage II, a tuft was formed by fusion of several filaments at their bases. Stage III represents the formation of a central rachis and development of serially fused barbs (III A) — to which, at a slightly later stage (III B), secondary barbs (barbules) were added. The two other stages, IV (bipinnate feathers with elaborate barbules and a closed vane) and V (the asymmetrical flight feathers of modern flying birds), are not shown. (Figure redrawn from ref. 5.)**

the Yixian Formation not only further strengthen the case for the theropod–bird connection, but also establish that feathers originated and initially diversified in non-flying non-avian theropod dinosaurs. Feathers pre-date the origin of birds and avian flight. They clearly evolved for some purpose other than flight, perhaps thermal insulation[3] or behavioural display[4] (or both) — functions they retain in present-day birds. ■

Hans-Dieter Sues is at the Royal Ontario Museum, 100 Queen's Park, Toronto, Ontario M5S 2C6, Canada.
e-mail: hdsues@rom.on.ca

1. Padian, K. & Chiappe, L. M. *Biol. Rev.* **73**, 1–42 (1998).
2. Feduccia, A. *The Origin and Evolution of Birds* 2nd edn (Yale Univ. Press, New Haven, 1999).
3. Regal, P. J. *Q. Rev. Biol.* **50**, 35–66 (1975).
4. Mayr, E. in *The Evolution of Life* (ed. Tax, S.) 349–380 (Univ. Chicago Press, 1960).
5. Xu, X., Zhou, Z.-H. & Prum, R. O. *Nature* **410**, 200–204 (2001).
6. Ji, Q., Norell, M. A., Gao, K.-Q., Ji, S.-A. & Ren, D. *Nature* **410**, 1084–1088 (2001).
7. Chen, P.-j., Dong, Z.-m. & Zhen, S.-n. *Nature* **391**, 147–152 (1998).
8. Ji, Q., Currie, P. J., Norell, M. A. & Ji, S.-A. *Nature* **393**, 753–761 (1998).
9. Xu, X., Wang, X.-L. & Wu, X.-C. *Nature* **401**, 262–266 (1999).
10. Jones, T. D., Farlow, J. O., Ruben, J. A., Henderson, D. M. & Hillenius, W. J. *Nature* **406**, 716–718 (2000).
11. Gauthier, J. A. *Mem. Calif. Acad. Sci.* **8**, 1–55 (1986).
12. Prum, R. O. *J. Exp. Zool.* **285**, 291–306 (1999).

Optics

Holograms of atoms

John Spence

X-ray crystallography can be used to produce three-dimensional images of atoms providing they are arranged periodically. For small atomic clusters or molecules, electron holography could provide even sharper images.

"That's the last potato I'll dig," said the young Ernest Rutherford, perhaps the greatest experimental physicist of the last century, on hearing, on his father's farm in New Zealand, that he'd won a scholarship to the University of Cambridge. And so began a century in which most Nobel prizes in physics and chemistry have been won by scientists trying to figure out the structure or arrangement of atoms by scattering things off them. Calculating structure from scattering patterns is known as 'inversion' — a problem Rutherford solved with help from a mathematician friend. Others who packed their bags for Stockholm included the team that developed methods for determining the structure of proteins from X-ray scattering, and Denis Gabor, who invented holography, a technique for creating three-dimensional images without using a lens.

As they describe in *Physical Review Letters*, a group of Swiss and Italian scientists[1] working at the Trieste X-ray synchrotron has now found a way to apply Gabor's method more effectively to electron scattering patterns, and so have produced strikingly beautiful three-dimensional holographic images of the local arrangement of atoms near the surface of a crystal. The technique Wider *et al.*[1] use is known as internal-source holography because it exploits an internal source of radiation to probe below the surface of structures. The hope is that this method can now be applied to reveal the structure of small atomic clusters that cannot be crystallized for analysis by conventional X-ray crystallography — from biological macromolecules to catalyst particles and nanostructures.

In Gabor's original proposal[2], a small

100 YEARS AGO

Dr. B. Sharp described, at a recent meeting of the Academy of Natural Sciences of Philadelphia, some observations he has made on the contents of the stomachs of the common cod. Several hundred stomachs were opened with the hope of finding shells of gastropods and bivalves. Numerous valuable shells were taken from the cod years ago by Stimpson and Gould on the New England coast, north of Cape Cod, and it was supposed that similar finds would come to light from the cod caught off Nantucket. The stomachs examined were filled almost exclusively with crustaceans and for the most part with species of *Panopeus*. Hermit crabs, without shells, and a few Crepidulae were also seen. Here and there young lobsters were found in the stomachs, occasionally two in one stomach. Dr. Sharp believes that the decrease in quantity of the lobsters, which has been so marked within the past few years, is partly due to their consumption by the cod; and as these have of late greatly increased in numbers, owing to the work of the United States Fish Commission, the lobsters have not been able to keep pace with the increase of their enemies.

From *Nature* 25 April 1901.

50 YEARS AGO

From the news in the Press it might be thought that ordinary administration and, more especially perhaps, forestry administration would be more or less at a standstill in Malaya. The annual report for 1949 of the Federal Forest Administration of the Federation of Malaya, by J. P. Edwards, acting director of forestry, shows the reverse, particularly in the professional and research sides of the work. It is true that the report states that, in connexion with saw milling, "Bandit activity has also caused the sawmills much difficulty. One was burnt down, several more have closed down to avoid being forced to provide supplies to the bandits, others have had lorries destroyed and many find difficulty in persuading their logging gangs to remain at work. The Police, too, are often obliged to clear the forest of loggers so that they are not mistaken for bandits during security operations. In spite of this millers have shown great determination to keep producing and develop their mills. Eleven new mills were planned and five others increased their plant."

From *Nature* 28 April 1951.

© 2001 Macmillan Magazines Ltd

external source of radiation illuminates a semi-transparent object, so that a magnified shadow-image (the hologram) appears on a distant screen, much as a child might project rabbit's ears on a wall by making finger-shapes in front of a lamp. Now replace the lamplight with the wavelike electron emitted by an atom just below the surface of an aluminium crystal. The electron can be liberated from the emitter atom by exciting it with an X-ray beam — the photoelectric effect. If you detect the photoelectrons emerging from the crystal surface on a hemispherical screen, then you have the essential arrangement of Wider *et al.*[1] (Fig. 1a). The image of the atoms in the surface layer (the scatterers) is projected by the emitter (source) atom in the layer below. (In fact, because the emitter emits, and the scatterers scatter, in all directions, one obtains a three-dimensional image of all the atoms around the emitter atom.)

Gabor found that the finest detail recoverable from his holograms was about equal to either the wavelength of the radiation or the size of the source — whichever was larger. Wider *et al.* use a wavelength of about 0.04 nanometres, whereas the photoelectron comes from an even smaller region inside an atom, so its wavelength could be shorter still. The small source size enables higher resolution, but makes the emitted electron waves highly uniform (coherent), so that the shadow-image becomes an uninterpretable interference pattern, whose inversion Gabor struggled with unsuccessfully. The idea of using an internal source can be traced back even further. In 1939, Bragg inverted X-ray scattering patterns from crystals containing strongly scattering internal 'source' atoms, by using visible light for the reconstruction[3] — a two-step process similar to Gabor's 1949 holography. More recently Bartell and Ritz[4] used an internal atomic source for electron holography of atoms in a gas.

Interest in photoelectron holography started in the late 1980s when Szöke[5] proposed the internal-source method used by Wider and colleagues[1]. A three-dimensional inversion scheme was introduced by Barton[6] for point-like scatterers, and the first experimental results were obtained by Harp *et al.*[7,8]. You may be wondering why every scattering atom doesn't also act as a source — it does, but because every atom lies among identical atomic surroundings in the periodic structure of the simplest crystals, it produces an identical hologram. For atomic clusters lying on a crystal surface, one can tune the incoming X-rays to excite one particular atomic species, which then alone acts as an emitter. In order for all the clusters to produce identical superimposed holograms, they must all lie in the same orientation, but need not be periodically arranged.

Physicists applying Barton's inversion scheme ran into several problems — notably a 'twin image' problem, which Gabor was famously unable to solve. The twin image is a ghostly inverted image, which appears superimposed on the true image in most inversion schemes. This problem and another involving multiple scattering have been addressed with some success over the past decade by adding together holograms recorded at several electron energies. A third problem arises because the emitting and scattering atoms do not emit pure spherical waves as the inversion schemes assume. Instead, scatterers scatter too strongly in the direction of the incident electron (forward scattering), especially at the higher energies (smaller wavelengths) needed for good resolution.

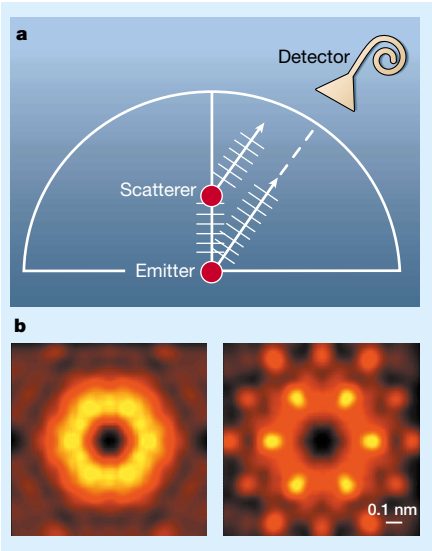


Figure 1 **Looking deeper with internal-source holography.** a, **In their experiment, Wider *et al.*[1] use X-rays to excite an emitter atom below the surface of an aluminium crystal. The emitted photoelectron can reach the detector either directly or by way of a scatterer atom in the surface layer. These two paths produce an interference pattern at the detector (similar to Young's fringes) which can be inverted to provide a three-dimensional image of the atoms around the emitter.** b, **Electron holograms of aluminium crystal (yellow shows highest intensity) from diffraction data collected at ELETTRA, Trieste[1]. The left hologram does not resolve atomic positions well because of forward scattering problems. The emitter atom is in the centre, and appears dark because it cannot image itself. Now, by controlling the angle between the detector and the emitted radiation, Wider *et al.* are able to distinguish nearest and next nearest positions of neighbouring atoms. The right hologram shows 18 atoms lying in one plane of an aluminium crystal. In fact, because of the three-dimensional nature of the holograms, Wider *et al.* can locate all 12 of the emitter atom's nearest neighbours (three above, three below and six in the same plane).**

The achievement of Wider *et al.* is to solve the forward-scattering problem, which they do by adjusting the polarization direction of the incident X-rays, so that the detector is always positioned close to an intensity minimum in the emitter's emission pattern. Practically speaking, the crystal then has to be rotated with respect to the synchrotron and the detector. The result is a more accurate inversion that clearly shows all the surrounding atoms out to much greater distances than previously possible, and in their correct places (Fig. 1b). Forward scattering is a problem when the emitter atom lies further from the detector than the scatterer, as in Fig. 1a. For the more interesting case of a cluster of foreign atoms on a crystal surface (using the top atom as the emitter) this problem doesn't arise, because the electrons must reverse direction to reach the detector. Nevertheless, some adsorbate emitters can lie inside a crystal surface, putting them beyond the scatterer.

These internal-source ideas have spawned other sub-fields[9,10]: X-ray fluorescence holography (XFH) and its time-reversed inverse, IXFH, and Auger fluorescence holography. X-ray fluorescence holograms are excited by synchrotron X-ray beams, and do not suffer from multiple scattering or uneven scattering problems, although the scattering of X-rays by atoms is much weaker (especially for light elements) than for electrons, and radiation damage effects differ. At the very least, these methods provide a starting point on which to base highly accurate calculations of the scattering patterns, and the internal-source method fills a niche between the extended X-ray absorption fine structure technique, which is used to characterize local atomic structure, and crystallography.

Researchers in Berlin, Brookhaven, Grenoble, Berkeley and elsewhere are beginning to apply internal-source methods to locate dopant atoms in semiconductors and alloys, to probe ordering in alloys, quasicrystals and magnetic atoms, and to look at buried interfaces, epitaxial films or macromolecules. Perhaps, with the next generation of brighter synchrotrons, internal-source holography may one day provide three-dimensional holographic movies of chemical processes in action on the atomic scale. If radiation damage doesn't prevent it, we may also see holograms of the membrane proteins, important for drug delivery, which are so difficult to crystallize for analysis by X-ray crystallography. The challenge now is to find ways to prepare arrays of similarly oriented clusters for the new internal-source holographies. ■

John Spence is in the Department of Physics and Astronomy, Arizona State University, Tempe, Arizona 85287-1504, USA.
e-mail: spence@asu.edu

1. Wider, J. *et al. Phys. Rev. Lett.* **86,** 2337–2340 (2001).
2. Gabor, D. *Nature* **161,** 777 (1948).

© 2001 Macmillan Magazines Ltd

3. Bragg, W. L. *Nature* **143,** 678–680 (1939).
4. Bartell, L. & Ritz, C. *Science* **185,** 1163–1165 (1974).
5. Szöke, A. in *Short Wavelength Coherent Radiation: Generation and Applications* (eds Attwood, D. T. & Boker, J.) 361–467 (AIP, New York, 1986).
6. Barton, J. J. *Phys. Rev. Lett.* **67,** 3106–3109 (1991).
7. Harp, G., Saldin, D. & Tonner, B. *Phys Rev. B* **42,** 9199–9202 (1990).
8. Fadley, C. S. *Prog. Surf. Sci.* **54,** 341–386 (1997).
9. Faigel, G. & Tegze, M. *Rep. Prog. Phys.* **62**, 355–392 (1999).
10. Cowley, J. M. *Acta Crystallogr.* **17,** 33–42 (1964).

Genome sequencing

A grin without a cat

Paul R. Gilson and Geoffrey I. McFadden

In some types of unicellular algae, the chloroplasts have their own nucleus — a legacy of the time when the chloroplast was a free-living cell. The sequence of the genome in one such nucleus is now revealed.

The electronics industry's mania for miniatures has brought us nifty gadgets such as pocket PCs, handheld televisions and wristwatch phones. But these devices are giants compared with what nature can produce. On page 1091 of this issue[1], Douglas and colleagues present an extraordinary case of genetic miniaturization — the genome sequence of an amazingly small nucleus. The nucleus in question is the so-called nucleomorph of a cryptomonad alga, *Guillardia theta*, and its genome weighs in at a mere 0.55 million base pairs. Compared with the human genome (3,200 million base pairs)[2,3], this nucleomorph sequence is sub-Lilliputian. How can two blueprints be so different?

In fairness, the nucleomorph is not a complete nucleus but a relic, its genome having been distilled to its essence by hundreds of millions of years of enslavement. Nucleomorphs are the highly reduced nuclei of 'endosymbiotic' algal cells that, in the distant past, set up home within unicellular hosts to mutual benefit. Like their hosts, the endosymbionts were eukaryotic, meaning, loosely, that they had a nucleus. Importantly, they were also photosynthetic, feeding their hosts with the products of this chemical reaction — carbohydrates and oxygen.

Such microscopic gardening arrangements are common. But the cryptomonad endosymbiosis belongs to a special category known as secondary endosymbiosis, whereby the photosynthetic captive becomes an integral, enduring part of the host cell. In the case of *G. theta*, the endosymbiont became what is known as a complex chloroplast. Over time — perhaps as many as 600 million years — the nucleus of this endosymbiont lost most of its genes. Nevertheless, Douglas *et al.*[1] show that the tiny vestige, the nucleomorph, is a bona fide nucleus. It has the usual eukaryotic trappings: several linear chromosomes; introns (sections of DNA that interrupt the coding sequence of genes); possible centromeres (the regions on eukaryotic chromosomes to which the chromosome-separating apparatus attaches during cell division); and histones (proteins that are swaddled by DNA). Indeed, the nucleomorph is a fairly typical nucleus but for two features — an impoverished complement of genes, and an almost complete lack of non-coding DNA.

Douglas *et al.* find that this cryptomonad nucleomorph has only 531 genes, whereas humans[2,3] boast at least 31,000. But it is gene density that makes the two genomes so different. Genes make up a mere 1% of the human genome[2,3]. The other 99% — often referred to as junk or non-functional DNA, which has largely unknown functions — may simply be the accumulated clutter of a system with slovenly housekeeping. Nucleomorphs, on the other hand, are the epitome of neatness and compactness. Many of the *G. theta* nucleomorph's genes have few or no spaces between them, and 44 genes even overlap, parsimoniously using both strands of the chromosomes, rather than the usual one[1]. The human genome seems profligate by comparison. Indeed, the entire nucleomorph genome would fit comfortably in one of the many yawning gaps between human genes.

Why are these two genomes so different? Evolutionary forces that shape genome size are not well understood. But we believe that streamlining of the nucleomorph genome is unlikely to be driven solely by natural selection for minimal DNA content. Rather, it may be a result of uncontrolled DNA loss, and it is here that comparison of these two genomes might be enlightening. Unlike other human chromosomes, the Y chromosome has no complementary partner, so it cannot undergo recombination during meiosis (the type of cell division that generates gametes, such as eggs and sperm). Recombination is a process by which mutations — particularly insertions and deletions of sequence — can be weeded out in subsequent generations[4]. The absence of recombination has resulted in wholesale loss of DNA from the Y chromosome, which is now a mere stump with only 50 million base pairs[2,3]. This chromosome persists, however, because it carries a handful of vital genes, in particular the testis-determining factors.

Nucleomorphs do not have sex chromosomes and each of the three chromosomes is thought to be paired[5]. Nevertheless, nucleomorph chromosomes are probably denied the normal opportunity to recombine because cryptomonad endosymbionts do not seem to undergo meiosis followed by genetic exchange with other nucleomorphs[1]. We believe that this genetic isolation and consequent lack of error-correcting mechanisms has caused nucleomorphs to slide into mutational hyperdrive and wholesale DNA loss. But, just as the Y chromosome is maintained because it produces something vital (such as testicles), so too must the nucleomorph genome endure — in this case, because it encodes components of the chloroplast[1].

The interesting question really is not 'How did the nucleomorph genome get so small?', but rather 'Why did it stop where it did?'. Just as the Cheshire Cat in Lewis Carroll's *Alice in Wonderland* faded until only its grin remained, nucleomorphs too appear to have reached an end-point. This applies not only to cryptomonads but also to chlorarachniophyte algae, the nucleomorphs of which also have three chromosomes and similarly small genomes[6]. The nucleomorphs of these two types of algae were derived from different secondary endosymbioses[1,7], so why have their genomes condensed to a similar end-point? Douglas *et al.*[1] offer an attractive explanation.

Gene sequences from the *G. theta* nucleomorph indicate that, as in other eukaryotes, the DNA is wrapped around histone proteins, forming 'chromatin'[1]. However, in contrast to other eukaryotes, nucleomorph chromatin apparently does not condense into higher-order structures during cell division[8]. Douglas *et al.* calculate that uncondensed nucleomorph chromosomes are only just short enough to fit inside a nucleomorph. If the nucleomorph DNA were packaged into fewer than three chromosomes, then those chromosomes would be too large to segregate during cell division. Conversely, if the DNA were separated onto more than three chromosomes, they might be too small to survive[9].

As bonsai versions of the nucleus, nucleomorphs provide important genetic and evolutionary lessons. The path taken by cryptomonads and chlorarachniophytes to obtain chloroplasts, namely by engulfing other algal cells, is well worn. Most phytoplankton — the algal backbone of aquatic food chains — also acquired their chloroplasts in this second-hand way[10]. But in these phytoplankton, all genes have been transferred to the host nucleus from the engulfed nuclei, which have been lost. Nucleomorphs are kept only while they encode something necessary for survival, probably proteins required to operate and maintain the chloroplast.

© 2001 Macmillan Magazines Ltd

Surprisingly, the nucleomorph[1] contains only 30 genes encoding proteins needed to operate the chloroplast, with most such proteins being encoded in the host nucleus. Why are these 30 genes still found in the nucleomorph? Perhaps gene transfer is incomplete, or perhaps (more likely) it is blocked in some way. There is insufficient evidence either way, but again the sequence of the chlorarachniophyte nucleomorph, which also encodes a mere handful of chloroplast proteins[11], might be informative. By determining the extent of overlap between these subsets of stranded genes, we will be able to formulate hypotheses about gene transfer from nucleomorphs, just as we have for mitochondria and simple chloroplasts[12–14]. ■

Paul R. Gilson is at the Centre for Cellular and Molecular Biology, School of Biological and Chemical Sciences, Deakin University, Victoria 3125, Australia.

Geoffrey I. McFadden is at the Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Victoria 3010, Australia.
e-mail: g.mcfadden@unimelb.edu.au

1. Douglas, S. *et al. Nature* **410,** 1091–1096 (2001).
2. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
3. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
4. Orr, H. & Kim, Y. *Genetics* **150,** 1693–1698 (1998).
5. Hansmann, P. & Eschbach, S. *Eur. J. Cell Biol.* **52,** 373–378 (1990).
6. McFadden, G. I., Gilson, P. R., Hofmann, C. J., Adcock, G. J. & Maier, U.-G. *Proc. Natl Acad. Sci. USA* **91,** 3690–3694 (1994).
7. Van de Peer, Y., Rensing, S. A., Maier, U.-G. & de Wachter, R. *Proc. Natl Acad. Sci. USA* **93,** 7732–7736 (1996).
8. McKerracher, L. & Gibbs, S. P. *Can. J. Bot.* **60,** 2440–2452 (1983).
9. Murray, A. & Szostak, J. *Annu. Rev. Cell Biol.* **1,** 289–315 (1985).
10. Delwiche, C. F. *Am. Nat.* **154,** S164–S177 (1999).
11. Gilson, P. R. & McFadden, G. I. *Proc. Natl Acad. Sci. USA* **93,** 7737–7742 (1996).
12. Gray, M., Burger, G. & Lang, B. *Science* **283,** 1476–1481 (1999).
13. Race, H., Herrmann, R. & Martin, W. *Trends Genet.* **15,** 364–370 (1999).
14. Palmer, J. *Science* **275,** 790–791 (1997).

Earth science

Shaken, not stirred

Michael Manga

It is difficult to obtain a dynamic picture of the Earth's mantle. A study involving geophysical observations with geochemical implications shows that compositionally distinct megablobs contribute to the ebb and flow.

The slow and continual deformation of the Earth's mantle is seen in the motion of tectonic plates at the surface. Flow in the mantle is driven by density variations, which are caused by changes in temperature and composition. Unfortunately, the velocity, pattern and driving forces of flow cannot be observed directly in the mantle, as they can in other dynamic parts of Earth such as the oceans and atmosphere. Rather, studies of the Earth's deep interior rely on indirect geochemical and geophysical measurements that are sensitive to the Earth's structure and dynamics.

On page 1049 of this issue, Forte and Mitrovica[1] combine a wide range of geophysical observations with mineral physics data — such as density and sound speed in minerals — to develop a model of global mantle flow and structure (Fig. 1a). They show that large-scale heterogeneity in the composition of the deep mantle may provide clues about the long-term evolution of the Earth.

The most tantalizing images of the Earth's interior have been obtained by studying the propagation of seismic waves generated by large earthquakes. Using a technique called seismic tomography — the geophysical analogue of a medical CAT scan — various properties of seismic waves, such as their travel times, can be used to infer the three-dimensional pattern of seismic-wave velocity in the Earth. It is tempting to translate wave propagation speeds into temperature variations in the mantle, and then to infer a model of flow. For example, waves will propagate slowly through regions that are hot, implying that such regions are less dense and will rise buoyantly. Conversely, faster waves imply a colder, denser region that will sink. Indeed, in the lower mantle the correlation of regions of fast waves with the expected locations of cold, subducted (sunken) oceanic plates (Fig. 1b) is usually used to argue that convection involves the entire depth of the mantle[2]. That is, the mantle cannot consist of distinct, isolated layers that do not mix by convection.

Deducing flow patterns from wave speeds can sometimes be perilous, however. For example, the large, fast anomalies of seismic velocity in the upper mantle below many continents are probably due to differences in composition — the 'tectosphere' in Fig. 1b — rather than regions of sinking mantle[3]. Recent models based on seismic tomography[4] also suggest that both compositional and thermal structures exist in the deep mantle[5].

Forte and Mitrovica[1] develop a global convection model based on mineral physics data — specifically, the dependence of density and seismic wave speeds on temperature and composition. Their model incorporates a large number of global geophysics observations, including gravity measurements, the motion of tectonic plates, and deformations at the Earth's surface and core–mantle boundary that are caused by flow. In short, it is the most comprehensive integrated study so far of mantle flow.

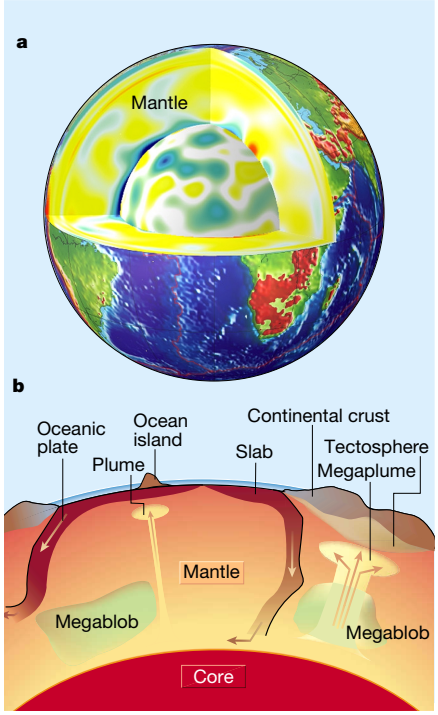


Figure 1 **The interior structure of the Earth.** a, **A cutaway of the Earth, showing temperature variations in the mantle and changes in composition at the core–mantle boundary. (Courtesy of Forte and Mitrovica.)** b, **The structure of the mantle, showing subducted slabs, the tectosphere and megablobs (discovered by Forte and Mitrovica[1]), which all influence temperature and flow (not to scale).**

Forte and Mitrovica verify that the whole mantle appears to act as a single convective system that is driven primarily by thermal anomalies. They also find evidence of large-scale compositional heterogeneity — megablobs — within the lower mantle. The magnitude of the density variations is sufficient to lead to the 'doming' mode of convection, in which large blobs of compositionally distinct mantle 'shake' — that is, the blobs move up and down every so often[6]. This finding is at odds with recent proposals that the deep mantle is convectively isolated from the rest of the mantle[7].

The results are not definitive, because geodynamic models are limited by the reliability and availability of data. The models of Forte and Mitrovica are even further limited by their use of simple assumptions about the composition of the lower mantle. But the new approach does provide a framework for incorporating improved seismic models and mineral physics data. In particular, much of the mineral physics data is extrapolated from measurements made at low pressures and temperatures, and will undoubtedly be refined.

The compositional variations and temperature anomalies mapped by Forte and

© 2001 Macmillan Magazines Ltd

Mitrovica, along with the mantle flow, are snapshots of a dynamic Earth. Such snapshots do not tell us how the megablobs developed, or how they will evolve. Are these blobs 'unstirred' regions that have persisted throughout most of the Earth's history? If so, can they be the physical manifestation of the mantle reservoir that still contains elements that usually tend to be concentrated in the crust? This mantle reservoir is thought to provide the distinctive isotopic characteristics of ocean island basalts formed by mantle plumes originating in the deep mantle[8] (Fig. 1b). The high viscosity of the lower mantle, inferred from the dominance of long-wavelength seismic structure[9], slows the rate of mixing. But numerical models of convection have shown that high viscosity in the lower mantle is not sufficient to prevent significant stirring of heterogeneities[10]. So, if the blobs are old, they must be much more viscous than the rest of the mantle[11]. Alternatively, the megablobs could be actively forming and growing, perhaps through the segregation and accumulation of subducted materials[12].

Solving these puzzles will require integrating fluid-dynamic models of thermochemical convection with an even wider range of observations, in particular those that depend on changes to the mantle. Because this evolution of the mantle is coupled to that of the core and the continental crust, determining the nature and origin of deep mantle structures may ultimately provide new insights into the coupled chemical and dynamic evolution of the entire Earth. ■

Michael Manga is in the Department of Earth and Planetary Sciences, University of California, Berkeley, California 94720, USA.
e-mail: manga@seismo.berkeley.edu

1. Forte, A. M. & Mitrovica, J. X. *Nature* **410,** 1049–1056 (2001).
2. Grand, S. P., van der Hilst, R. D. & Widiyantaro, S. *GSA Today* **7,** 1–7 (1997).
3. Jordan, T. H. *Rev. Geophys.* **13,** 1–12 (1975).
4. Saltzer, R. L., van der Hilst, R. D. & Karason, H. *Geophys. Res. Lett.* **28,** 1335–1338 (2001).
5. Stacey, F. D. *Phys. Earth Planet. Int.* **106,** 219–236 (1998).
6. Davaille, A. *Nature* **402,** 756–760 (1999).
7. Kellogg, L. H., Hager, B. H. & van der Hilst, R. D. *Science* **283,** 1881–1884 (1999).
8. Hofmann, A. W. *Nature* **385,** 219–229 (1997).
9. Bunge, H.-P. & Richards, M. A. *Geophys. Res. Lett.* **23,** 2987–2990 (1996).
10. van Keken, P. & Zhong, S. *Earth Planet. Sci. Lett.* **171,** 533–547 (1999).
11. Manga, M. *Geophys. Res. Lett.* **23,** 403–406 (1996).
12. Christensen, U. R. & Hofman, A. *J. Geophys. Res.* **95,** 19867–19884 (1994).

Cancer

Improved mouse models

Anton Berns

Mice that have been genetically modified to be prone to tumours are thought to be poor models of sporadic cancer. This problem can be tackled with a new generation of modified mice.

Cells tend to become cancerous only after they have suffered mutations in several genes, including those that control the mechanisms by which cells multiply, die or migrate to other parts of the body[1]. This requirement for several specific genetic alterations poses a barrier to cancer, not least because mutations occur more-or-less at random throughout the genome. Nevertheless, a single, widespread cancer-promoting mutation can dramatically increase the occurrence of tumours. Such a situation occurs for hereditary tumours[2], as well as in mice that have been genetically engineered to provide models for studying cancer[3]. This is one of the reasons why such mice are generally poor representations of randomly arising (sporadic) tumours. By using a cunning genetic technique, however, Johnson and colleagues have produced a much more convincing mouse model of sporadic human lung cancers, as they report on page 1111 of this issue[4].

Given that the accumulation of mutations is thought to be required for tumour formation, it seems strange that, in current mouse models or in hereditary cancers, one widespread genetic alteration can confer such a strong predisposition to cancer. A fundamental difference between inherited and sporadic cancers may lie at the heart of this conundrum. A cell that is accumulating sporadic tumour-promoting mutations is, at least at first, surrounded by normal cells. These can suppress the propensity of the mutant cell to proliferate. By contrast, when someone has inherited a cancer-promoting mutation, that mutation would be present in many of their cells — a group of such mutant cells would enhance its own proliferation.

According to this concept, a group or 'field' of mutant cells is a tumour-promoting condition in its own right. The observation of 'field cancerization' in cancer patients provides support for this idea[5]. *In vitro*, meanwhile, adding the *Ras* oncogene (a generic name given to three mutant genes that can accelerate cell proliferation) can cause cells to become cancerous, but only in the absence of normal cells[6]. So one of the steps towards sporadic cancer might be the development of a field of mutant cells[7]. Classical mouse models do not incorporate this transition from a single mutant cell to a field of cells, because fields already exist. New models are therefore urgently needed, and Johnson *et al.*'s genetically modified mice[4] may meet this demand.

Using intricate genetic techniques, the authors have produced a strain of mice in which an active, mutant *K-ras* gene is generated at random in some cells by means of spontaneous recombination — the shuffling of segments of DNA — between or within chromosomes (Fig. 1, overleaf). This randomness echoes the sporadic occurrence of *K-ras* mutations in many human cancers.

The model has several unique features. First, *in vivo* recombination creates a low but substantial incidence of the active *K-ras* oncogene (the authors estimate that it arises once every 10^3 to 10^7 cell divisions). This gives rise to scattered cells that express that gene. Second, recombination yields an active *K-ras* oncogene within the authentic genetic locus on the correct chromosome. So, at least initially, expression of the mutant gene is under normal physiological control. Third, as the switch occurs randomly, all types of tissue, throughout development, can theoretically experience the effects of the mutant gene. Finally, the mutant gene can be traced easily in fields of cells by standard molecular biological techniques.

The mice showed a high incidence of lung cancers — including adenocarcinomas, which are predominant in humans and have a particularly poor prognosis — as well as some lymphomas (cancers of the blood) and skin papillomas. Metastasis, the migration of tumour cells to other parts of the body, occurred infrequently. The authors then produced mice that also had a deficiency in p53 — a protein that normally inhibits cancer development. These doubly mutant mice had a greater number of malignant tumours.

Analysis of groups of cells from mice with only the *K-ras* oncogene showed that this gene was activated in small foci (fields) of tumour cells — an early feature of cancer — but not in flanking regions (T. Jacks, personal communication). However, it is not clear whether the foci arose from a single cell with mutant *K-ras*, in other words, whether mutant *K-ras* drives the formation of a focus from a single mutant cell that is surrounded by normal cells. It is possible that the recombination event occurred in a single cell before birth, when *K-ras* (whether mutant or normal) is not yet switched on[8]. Such a cell would multiply as usual during development, forming a field of cells with a silent *K-ras* oncogene. After birth, that field of cells would suddenly start to express *K-ras*, and this might drive the formation of foci. Moreover, given that several distinct stages of lung cancer, for example, were seen in the mutant mice[4], further genetic alterations are no doubt involved in tumour progression. It will be interesting to identify these changes.

Although mutations in the human *K-ras* gene are common in pancreatic and colon cancers, the mutant mice did not have these

© 2001 Macmillan Magazines Ltd

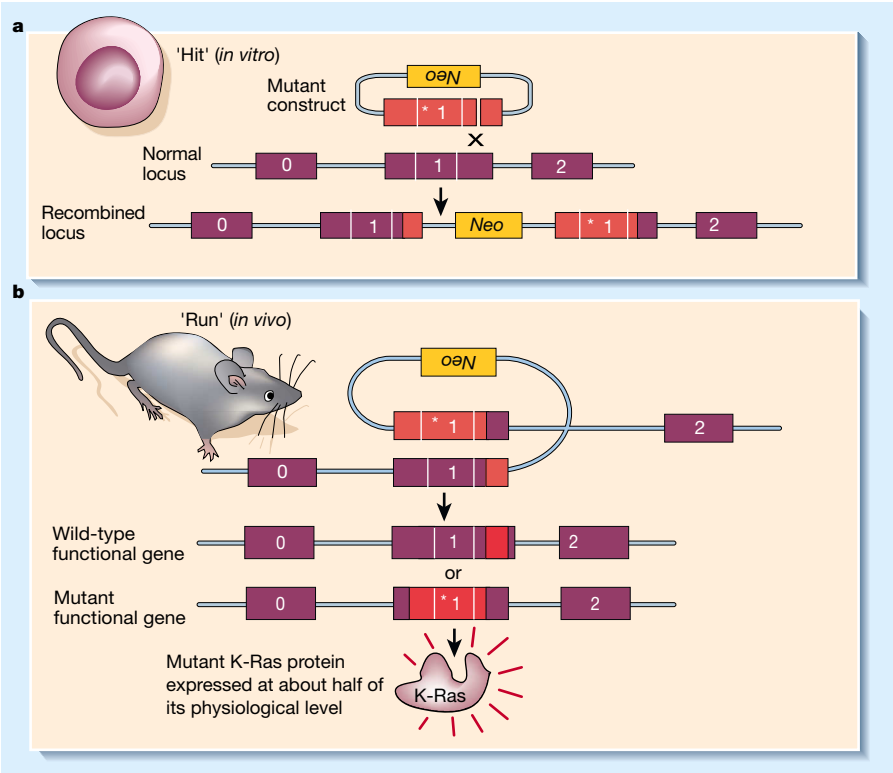


Figure 1 **A new 'hit-and-run' strategy for generating mouse models of sporadic cancers[4].** a, **A 'silent' mutant *K-ras* gene (oncogene) is inserted into the site (locus) of the authentic, normal *K-ras* gene in mouse embryonic stem cells, which are taken from early embryos. The insertion involves the mutant and normal DNA segments being shuffled (recombined). The altered stem cells are then used to generate mice. The asterisk represents the mutation. *Neo* is the neomycin gene, which allows cells that bear it to grow when cultured with the antibiotic neomycin. This enables researchers to identify cells that have taken up both this gene and the adjoining *K-ras* oncogene. 0, 1 and 2 represent the protein-coding sections of the *K-ras* gene.** b, **Within the mice, spontaneous recombination occurs either within or between chromosomes during cell division. Occasionally, this will result in the *Neo* gene being excised from the *K-ras* locus, producing a functional wild-type or mutant *K-ras* gene in random tissues. The active mutant K-Ras protein drives tumour growth.**

tumours. This could be because recombination frequencies are low in these tissues, or because different cells in different species respond differently to mutant *K-ras*. Indeed, cell-type-specific effects are almost certainly involved, as foci of early tumour cells in the animals' intestines did not become fully cancerous. Intriguingly, though, when the authors produced mice bearing both the *K-ras* oncogene and Apc^{min} — a mutated gene that predisposes humans to intestinal adenomas — the type and frequency of intestinal tumours did not change. One wonders whether the *K-ras* mutant was activated in tumours in these doubly mutant mice.

Jackson *et al.*'s mutant mice[4] represent a unique new model. Others, using different recombination techniques[9] to produce sporadic mutations, will no doubt follow suit. Sporadic activation of a recombination system *in vivo* has another advantage: it allows researchers to introduce several genetic alterations at a defined moment in a subset of cells, increasing the ability to mimic a wide spectrum of human tumours. These models should allow us to work out the sequence of events, including field cancerization, in tumour development. They also seem well suited for testing prevention strategies and treatments. A further practical refinement of these models will undoubtedly include sensitive *in vivo* imaging. For example, it might prove possible to activate luciferase, a light-generating enzyme, only in specific tumours — a promising way by which such features as tumour growth, spread, regression and relapse can be monitored non-invasively[10]. Expectations are high, but the proof of the pudding will be in the eating. ■

Anton Berns is in the Division of Molecular Genetics, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands.
e-mail: tberns@nki.nl

1. Hanahan, D. & Weinberg, R. A. *Cell* **100,** 57–70 (2000).
2. Kinzler, K. W. & Vogelstein, B. *Cell* **87,** 159–170 (1996).
3. Tuveson, D. A. & Jacks, T. *Oncogene* **18,** 5318–5324 (1999).
4. Johnson, L. *et al. Nature* **410,** 1111–1116 (2001).
5. Garcia, S. B., Park, H. S., Novelli, M. & Wright, N. A. *J. Pathol.* **187,** 61–81 (1999).
6. Land, H., Chen, A. C., Morgenstern, J. P., Parada, L. F. & Weinberg, R. A. *Mol. Cell. Biol.* **6,** 1917–1925 (1986).
7. Deng, G., Lu, Y., Zlotnikov, G., Thor, A. D. & Smith, H. S. *Science* **274,** 2057–2059 (1996).
8. Thrane, E. V. *et al. Exp. Lung Res.* **23,** 35–49 (1997).
9. Sauer, B. *Methods* **14,** 381–392 (1998).
10. Contag, C. H., Jenkins, D., Contag, P. R. & Negrin, R. S. *Neoplasia* **2,** 41–52 (2000).

Daedalus

New life

The British House of Lords has ruled that stem cells from fetuses can be used in medical research; nonetheless, we would all like a way of minimizing this source. Stem cells, of course, are only a few generations away from their creation. They express on their surface few or none of the histocompatibility molecules which bring the immune system down on them so heavily. Furthermore, the brain is a relatively privileged immunological site. That is why victims of Parkinson's disease, for example, can receive an injection of alien cells which will not be rejected. The cells may take their developmental cues from the brain around them, and reverse the effects of the disease.

So, says Daedalus, the best way forward would be to multiply stem cells *in vitro*, so that they could all be generated from just one fetus. DREADCO biochemists think that the medium is important. The best medium would be very 'young' — it should contain spinal fluid, and so forth, from very young organisms. It should be devoid of polarity cues, as this would encourage the cells to remain pluripotent and would repress any development or the establishment of gradients. A stirred system should do this.

So DREADCO biochemists are devising new media in which stem cells can be grown without deviating towards development. They are doping purely synthetic media with extremely young spinal fluid, brain fluid and so on, hoping that the cells will remain undeveloped for many generations. With luck these cells will be largely immune from immunological rejection. Thus in the brain they would take their developmental cues purely from the organ around them.

Indeed, says Daedalus, many simpler organisms, such as lizards, can regenerate legs or tails that become damaged or missing. They seem to be able to reverse cell development, and create new stem cells for the missing organs. With luck, plentiful human stem cells will enable humans to repair broken or missing arms and legs and wrecked tissues without the current trouble and limitations. All sorts of human damage could then be neatly repaired by allowing the cells to carry out the developmental plans which must be present even in the stump of a limb. Only the immune system, more active in the body than the brain, will have to controlled. And heart-transplant teams have already shown how to do this.

David Jones

© 2001 Macmillan Magazines Ltd

Polio vaccine samples not linked to AIDS

A search through the archives clears early vaccines of starting the AIDS pandemic.

It has been suggested that chimpanzee kidney cultures may have been used in the preparation of oral polio vaccine stocks used in Africa during the late 1950s, and so could have introduced the primate precursor of the immunodeficiency virus HIV-1 into humans[1,2]. Here we analyse frozen samples of the suspect vaccine by using the polymerase chain reaction (PCR) to amplify any HIV-1-related nucleic acids or chimpanzee mitochondrial DNA that might be present, but we have failed to detect either. Our findings do not support the hypothesis that HIV-1 was introduced by oral vaccination against poliovirus .

A short description of five samples of oral polio vaccine (OPV) held by the Wistar Institute is given in Table 1. CHAT pool 13 was widely used to vaccinate thousands of people (mainly children) in Léopoldville[3] in the former Belgian Congo (now Kinshasa, Democratic Republic of Congo) and some in Poland. Three further samples of CHAT vaccine held at the Centers for Disease Control (CDC) for 20–40 years were included in the study. The combinations of oligonucleotide primers and probes used can detect a wide variety of HIVs, including HIV-1 M, O and N isolates, as well as chimpanzee isolates Gab1, US and ANT70. The PCR assays have a sensitivity of about 50–80 RNA copies per ml. No sample proved positive for HIV-1-related nucleic acids by any of four primer pairs we used. By contrast, all samples were positive for poliovirus using PCR with reverse transcription, demonstrating that there was not extensive degradation of viral RNA (Table 1).

To identify the source of the kidney epithelial monolayers used to culture OPV, we analysed a small region (140 base pairs) within the mitochondrial 12S ribosomal RNA gene[4]. The primers we used amplify DNA from animals, including the African green monkey, chimpanzee, rhesus monkey and sooty mangabey (see supplementary information). All but one of the OPV samples were positive for mitochondrial (mt) DNA. With the exception of CHAT 1FL (see below), most sequences (86%) were highly related to those of the rhesus (*Macaca mulatta*) or cynomolgus monkey (*Macaca fascicularis*), and differed by only 0–2 bases from a known haplotype (Fig. 1).

A few sequences were nuclear paralogues of mtDNA. The amplification of bona fide mtDNA sequences also picks up these nuclear insertions of mitochondrial sequences (NUMTs). They behave as pseudogenes and fix mutations more slowly than mtDNA itself[5]. A few of these NUMTs, three out of more than 250 sequences, were related to mtDNA sequences of a variety of *Cercopithecus* monkeys. These *Cercopithecus*-like sequences never exactly corresponded with a known sequence in the databases, but differed by about 4 base pairs. Most rhesus monkey mtDNA sequences corresponded to known mtDNA haplotypes. Some of the *Cercopithecus*-like sequences were also derived from single rhesus monkey samples (not shown). We conclude that these really are NUMTs and not evidence that some African monkey

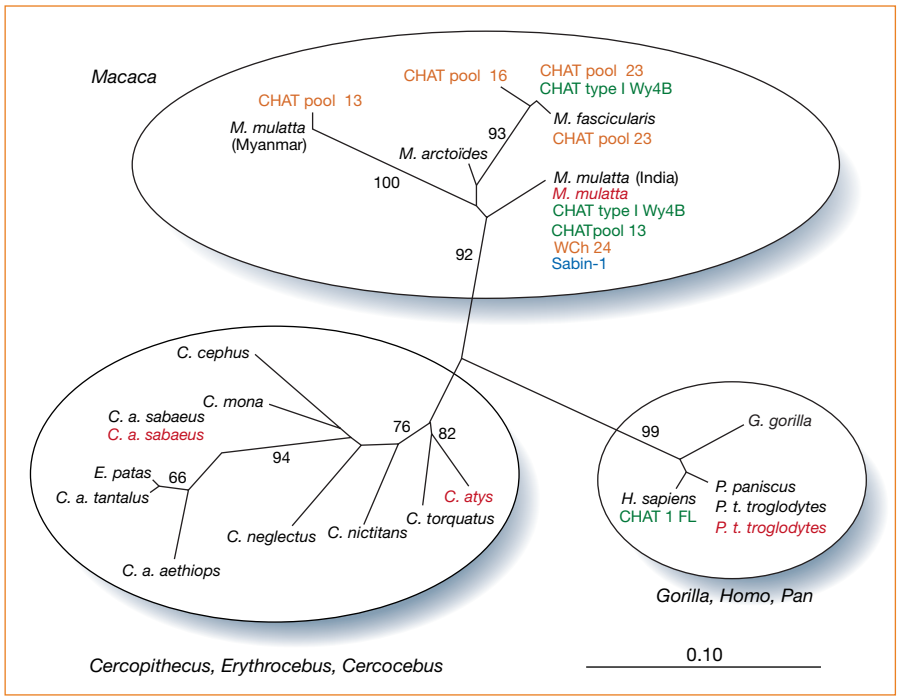


Figure 1 Neighbour-joining tree of 12S mtDNA sequences (see ftp.pasteur.fr/pub/retromol/wistar01). The three groups correspond to the *Macaca*, *Gorilla/Homo/Pan* and *Cercopithicus/Erythrocebus/Cercocebus* lineages. Of the latter, *C. atys* and *C. torquatus* are *Cercocebus* lineage. Wistar and CDC OPV samples and mtDNA controls are in yellow, green and red, respectively; the Sabin-1 positive control is in blue. Other sequences are from databases. The phylogenetic tree was derived by the neighbour-joining method applied to pairwise sequence distances calculated using the Kimura two-parameter method transition/transversion ratio set to 10. Branch lengths are drawn to scale: bar, 0.1 nucleotide replacements per site. Numbers indicate the percentage of bootstrap samples in which the cluster is supported (only bootstrap values over 60% are given). Rhesus macaques are split into two groups, corresponding to the Chinese and Indian subspecies. Remaining sequences are nuclear insertions of mtDNA sequences (see supplementary information).

Table 1 Oral poliovirus vaccine samples tested

Wistar samples		SIV_{CPZ}/HIV-1	Polio
CHAT pool 13	75,000 vaccinated in Léopoldville	–	+
CHAT pool 16 A-5	Closest passage to pool 13 available	–	+
CHAT pool 23 7.7. logs	Possibly used in other large trials; protocol on file	–	+
W Ch 24 57C-40 137-71	Possibly used in other large trials	–	+
W Ch 25	Late-passage virus	–	+
CDC samples			
CHAT pool 13	Monkey-kidney passage of CHAT pool 13 (29 Aug 1960)	–	+
CHAT type I Wy4B-5	Obtained from Wistar, made at Wyeth	–	+
CHAT 1FL	Late passage of CHAT 13 in human FL cells (15 Oct 1979)	–	+

SIV_{CPZ} is the isogenic chimpanzee counterpart of HIV-1 strains. For detection of SIV_{CPZ}/HIV-1 sequences, we used four primer pairs and two for poliovirus. A committee set up by the Wistar Institute identified frozen OPV samples that had been used in Central Africa or had been prepared during that era. Samples and controls were coded and hand-delivered to Roche Molecular Systems and the Institut Pasteur, each ignorant of the other's identity. Choice of amplification primers was left to the individual investigators to increase the chances of detection by at least one group. Small fragments were amplified in case there had been degradation of nucleic acids, particularly RNA, after being frozen for 40 years. Samples were also tested for poliovirus sequences (because if an OPV sample tested negative, then a negative result for SIV/HIV-1 and mtDNA would be meaningless). Infectious poliovirus was recovered for CHAT pool 16 and all three CDC-held samples (V. Racaniello, personal communication). A recent sample of Sabin-1 poliovirus grown on a rhesus monkey kidney cell line[11] was included as a positive control. mtDNA from four animals and a titration of HIV-1 isolate VQA, SIV_{CPZ} Gab1, HIV-2 NIH-Z stocks served as controls.

© 2001 Macmillan Magazines Ltd

kidneys, along with those of macaques, had been used to grow OPV.

Given that OPV was prepared from pools of culture supernatants derived from individual pairs of kidneys, the presence of multiple mitochondrial haplotypes in the same sample is not unexpected. This would also explain the 3:2 mix of *M. mulatta* and *M. fascicularis* haplotypes in CHAT type 1 Wy4B. The detection of *Homo* sequences in CHAT 1FL stems from the fact that the virus was grown on the FL human diploid cell line.

To confirm these findings, we studied a mitochondrial 12S-rRNA locus (101 base pairs) just 5′ to the previous segment. This too was highly informative, with ten point substitutions and one insertion/deletion distinguishing macaque and chimpanzee sequences. All of our findings were confirmed. As the Wistar CHAT pool-13 lot was widely used in Léopoldville, we increased the resolution of the analysis. No chimpanzee sequences were found at a resolution of 1.7%. As this frequency is below the reciprocal of the number of animals generally used to make a lot of OPV, we conclude that none of these OPVs were prepared using chimpanzee kidneys.

In addition, none of the samples was positive when amplified with primers specific for chimpanzee nuclear DNA (data not shown). The extensive use of rhesus and cynomolgus monkeys in the preparation of all the Wistar samples described here not only confirms earlier claims[6–8], as well as those denying the use of chimpanzee kidneys in the preparation of OPV[9], but also reflects the procedures used in the late 1950s by Salk and Sabin.

CHAT pool 13 was used in Léopoldville between August 1958 and April 1960. The earliest bona fide HIV-1 sequence was derived from a 1959 serum sample taken from an adult living in the suburbs of Léopoldville[10]. According to the OPV/AIDS hypothesis[1,2], the CHAT pool-13 sample should contain chimpanzee DNA, but as only macaque sequences were found, this tight geographical and temporal association would seem to have been a coincidence.

There is a corollary to our conclusion that the Wistar OPV samples were prepared using macaque kidneys. Given that simian immunodeficiency viruses (SIVs) are confined to African primates, the kidneys could not have been positive for SIV. In other words, the absence of SIV nucleic acids in all samples do not represent false negatives. We find no evidence to support the OPV/AIDS hypothesis.

Philippe Blancou*, Jean-Pierre Vartanian*, Cindy Christopherson†, Nicole Chenciner*, Claudio Basilico‡, Shirley Kwok†, Simon Wain-Hobson*

* *Unité de Rétrovirologie Moléculaire, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris cedex 15, France*
e-mail: simon@pasteur.fr
† *Department of Infectious Diseases, Roche Molecular Systems, 1145 Atlantic Avenue, Alameda, California 94501, USA*
‡ *Department of Microbiology, New York University School of Medicine, 550 First Avenue, New York, New York 10016, USA*

1. Curtis, T. in *Rolling Stone* 54–60 (1992).
2. Hooper, E. *The River: A Journey to the Source of HIV and AIDS* (Penguin, London, 1999).
3. Plotkin, S. A., Lebrun, A., Courtois, G. & Koprowski, H. *Bull. World Health Org.* **24,** 785–792 (1961).
4. van der Kuyl, A. C., Kuiken, C. L., Dekker, J. T. & Goudsmit, J. *J. Mol. Evol.* **40,** 173–180 (1995).
5. Saitou, N. & Ueda, S. *Mol. Biol. Evol.* **11,** 504–512 (1994).
6. Koprowski, H. *J. Am. Med. Assoc.* **178,** 1151–1155 (1961).
7. Plotkin, S. A. in *Proc. Sixth Intl Congr. Microbiol. Standardisation* 48–73 (Hoffman, Berlin, 1961).
8. Plotkin, S. A., Katz, M., Brown, R. E. & Pagano, J. S. *Am. J. Dis. Child.* **111,** 27–30 (1966).
9. Plotkin, S. A. & Koprowski, H. *Science* **286,** 2450 (1999).
10. Zhu, T. *et al. Nature* **391,** 531–532 (1998).
11. Hull, R. N., Cherry, W. R. & Tritch, O. J. *J. Exp. Med.* **165,** 903–917 (1962).

Supplementary information is available on *Nature*'s website at www.nature.com or as paper copy from the London editorial office of *Nature*. All sequences are available at ftp.pasteur.fr/pub/retromol/wistar01.

Vaccine safety

Analysis of oral polio vaccine CHAT stocks

Batches of experimental oral vaccines against poliovirus (OPV CHAT) that were administered in Central Africa in the 1950s have been implicated in the origin of the AIDS pandemic because of possible retroviral contamination during the vaccine's manufacture, which allegedly involved chimpanzee kidney cells[1]. Here we use a molecular analysis to show that two CHAT type-1 polio vaccine stocks were prepared from macaque and not chimpanzee cells, and contain neither human nor simian immunodeficiency virus sequences. Our results do not support the hypothesis that these materials were responsible for the transmission of the AIDS virus to humans.

The potential for oral polio vaccines to initiate the AIDS pandemic has been investigated previously[2,3]. Many species of African non-human primates are naturally infected with simian immunodeficiency viruses (SIV) and the common chimpanzee (*Pan troglodytes*) harbours SIV_{CPZ}, the closest relative to modern strains of the human immunodefiency virus HIV-1 (ref. 4). The use of chimpanzee cells to prepare CHAT vaccine may therefore have resulted in the inadvertent transmission of SIV_{CPZ} to humans.

To test this, we analysed two stocks of OPV CHAT. The first, labelled 19 CHAT 10A-11, was received by NIBSC in 1981 from the Karolinska Institute, Stockholm, and represents an original vial sent from the Wistar Institute, Philadelphia in 1958. The second, CHAT 6039 (an internal NIBSC number) made at the Institute of Immunology, Zagreb, was received by NIBSC in 1987. The presence of CHAT virus was confirmed by molecular assay and both samples yielded viable poliovirus in culture as evidence of quality of long-term storage (J. Martin, personal communication).

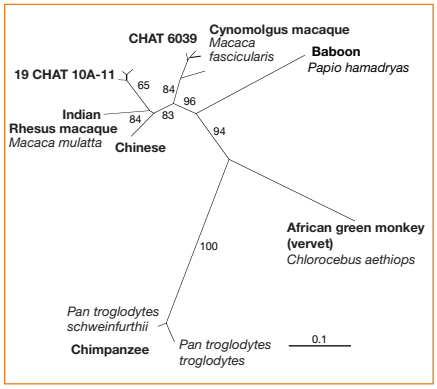


Figure 1 Nucleic-acid sequences were aligned using ClustalW (version 1.5) and evaluated using the Phylip Software package. Evolutionary distances were estimated using DNADIST (default: Kimura 2-parameter method, transition:transversion ratio 2.0) and phylogenetic relationships estimated by the neighbour-joining method using NEIGHBOR. Reproducibility of the branching pattern was assessed with SEQBOOT (boot strap, 100 replicates).

We tested extracted nucleic acid for HIV-1 RNA using polymerase chain reaction with reverse transcription (RT-PCR) assays based on *gag* (Amplicor HIV-1 Monitor assay, version 1.5) and combinations of long terminal repeat (LTR) sequences[5], which are capable of detecting HIV-1 group M (subtypes A–H)[6], group N and O viruses and SIV_{CPZ}. Additional combinations of LTR sequences used in nested PCR reactions were 507–529 and 524–543 (forward orientation), 622–641 and 628–648 (reverse orientation); the numbering is based on HIV-1 HXB2. HIV-1 primers in *pol* and *env* (gp41; ref. 7) and for HIV-2/SIV_{SM} (ref. 5) were also included. Sensitivity was shown to be less than 400 RNA equivalents per ml for each primer set. Using assays that detect this wide range of genetic variants of HIV-1, including SIV_{CPZ} and the earliest known sequence of HIV-1 present in the Belgian Congo in 1959 (ref. 7), we were unable to find any evidence of HIV/SIV in either CHAT stock.

Substrate cells were identified by targeting the D-loop control region of mitochondrial DNA using PCR. Chimpanzee-specific primers failed to amplify a product from either sample. Furthermore, these primers were capable of detecting limiting amounts of chimpanzee template in the presence of 10^5 macaque-cell equivalents, making it unlikely that chimpanzee cells were used to propagate poliovirus in the two prior passages of vaccine material. However, generic D-loop primers capable of amplifying DNA from baboon, African green monkey

© 2001 Macmillan Magazines Ltd

(vervet), chimpanzee, Chinese and Indian rhesus or cynomolgus macaques successfully yielded product from both batches. Sequencing and phylogenetic analyses indicate that the cells used to prepare 19 CHAT 10A-11 were obtained from rhesus macaques, and those for CHAT 6039 were from cynomolgus macaques (Fig. 1).

Failure to detect HIV/SIV sequences or chimpanzee cellular components in two OPV CHAT stocks, together with the positive identification of macaque mitochondrial sequences, provides no support for the hypothesis that these materials were responsible for the entry of HIV into humans and the source of AIDS.

N. Berry*, C. Davis*, A. Jenkins*, D. Wood†, P. Minor†, G. Schild‡, M. Bottiger§, H. Holmes*, N. Almond*

*Divisions of * Retrovirology and † Virology, ‡National Institute for Biological Standards and Control, Blanche Lane, South Mimms, Potters Bar, Hertfordshire EN6 3QG, UK*

e-mail: nberry@nibsc.ac.uk

§Karolinska Institute, 17177 Stockholm, Sweden

1. Hooper, E. *The River: A Journey to the Source of HIV and AIDS* (Penguin, London, 1999).
2. Garrett, A. J., Dunham, A. & Wood, D. *The Lancet* **342,** 932–933 (1993).
3. Dickson, D. *Nature* **407,** 117 (2000).
4. Hahn, B., Shaw, G., DeCock, K. & Sharp, P. *Science* **287,** 607–614 (2000).
5. Berry, N. J. *et al. J. Hum. Virol.* **1,** 457–468 (1998).
6. WHO Network for HIV Isolation and Characterisation *AIDS Res. Hum. Retrovir.* **10,** 1327–1343 (1994).
7. Zhu, T. *et al. Nature* **391,** 594–597 (1998).

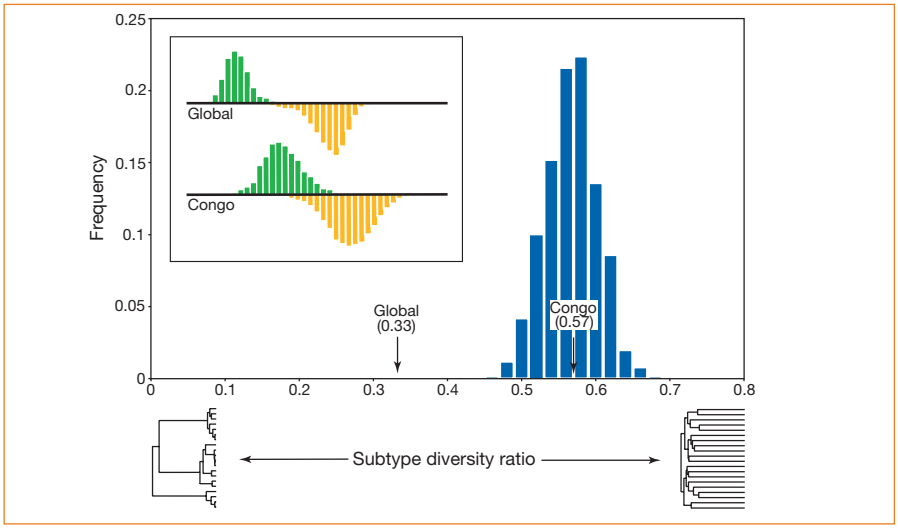


Figure 1 Maximum-likelihood phylogenies were estimated for three V3–V5 data sets: HIV-1 sequences from the Democratic Republic of Congo (423 base pairs), global isolates (426 base pairs), and Congo and global isolates combined (396 base pairs). Given a phylogeny with tips labelled according to subtype, the subtype diversity ratio (SDR) was calculated as the mean path length between tips of the same subtype divided by the mean path length between tips of different subtypes. For the phylogeny of the global isolates, 11 subtypes were allocated according to standard HIV-1 nomenclature[7]. For the Congo phylogeny, 11 subtypes were allocated so as to minimize the SDR score, using a heuristic optimization algorithm. This assignment is that which gives the maximum possible subtype structure for the Congo phylogeny. The global phylogeny gave an SDR of 0.33 and the Congo a value of 0.57. The analysis was repeated after removal of the Congo and global sequences previously identified as intersubtype recombinants[4,5]. Our analysis will only be affected if recombination breakpoints fall within the V3–V5 region, so excluding recombinants changes the SDR only marginally (0.35 for the global phylogeny; 0.58 for the Congo). SDR values were similar when Congo isolates were assigned to different numbers of subtypes (for example, 0.59 and 0.55 in the case of 8 and 14 subtypes, respectively). To assess the significance of the difference between the global and Congo SDRs, we obtained a null distribution by simulating phylogenies under an exponential growth coalescent process inferred from *env* gene sequences of subtype A (ref. 6), which is common in Africa. The frequency distribution of minimum SDR values for these simulated phylogenies is shown in blue. Inset: normalized frequency distributions of intrasubtype path lengths (above the line) and intersubtype path lengths (below the line), plotted on the same horizontal scale (0.0–0.8 substitutions per site), for the global and Congo phylogenies. See supplementary information for details of trees and phylogenetic methods.

Human immunodeficiency virus

Phylogeny and the origin of HIV-1

The origin of human immunodeficiency virus type 1 (HIV-1) is controversial. We show here that viruses obtained from the Democratic Republic of Congo in Africa have a quantitatively different phylogenetic tree structure from those sampled in other parts of the world. This indicates that the structure of HIV-1 phylogenies is the result of epidemiological processes acting within human populations alone, and is not due to multiple cross-species transmission initiated by oral polio vaccination.

According to the oral polio vaccination (OPV) hypothesis, the main (M) group of HIV-1 (the viruses responsible for the majority of global AIDS cases) emerged as a result of the vaccination of about one million people, who were largely living in the Congo from 1957–60, with an oral vaccine against polio virus that had allegedly been cultured in chimpanzee kidneys[1]. This is claimed to have enabled the transfer to humans of chimpanzee simian immunodeficiency virus, the closest relative of HIV-1.

Conversely, phylogenetic analysis of HIV-1 sequences indicates that group M originated before the vaccination campaign[2], supporting a model of 'natural transfer' from chimpanzees to humans[3]. If this timescale is correct, then the OPV theory remains a viable hypothesis of HIV-1 origins only if the subtypes of group M differentiated in chimpanzees before their transmission to humans.

It has been suggested that the distinctive structure of the global group-M tree, which has been called a 'starburst' because of the apparently simultaneous appearance of viral subtypes, is consistent with the transfer of multiple viral lineages from chimpanzees to humans[1]. To test this, we analysed partial *env* sequences (V3–V5) of 197 HIV-1 isolates sampled in 1997 from the Congo[4], a likely location for the origin of HIV-1 group M under both hypotheses.

A phylogeny comprising the Congo data, plus 223 sequences representing the global diversity of HIV-1 (including all known subtypes), reveals comparable genetic diversity in the Congo strains to that among global strains, with many Congo lineages falling basal to the origin of each subtype as currently defined by the phylogeny of global strains[5]. We tested whether the structure of the Congo phylogeny differed from that of the global HIV-1 M group by comparing the subtype diversity ratio (SDR) of the two phylogenies (Fig.1). This is defined as the ratio of the mean within-subtype pairwise distance to the mean between-subtype pairwise distance.

Rather than assigning the Congo isolates to subtypes by their phylogenetic relationship to global strains, we used a heuristic algorithm to assign subtypes such that the subtype diversity ratio was minimized. The Congo and global phylogenies differ significantly in the SDR statistic, with the former showing no more subtype structure than phylogenetic trees simulated under a model of exponential population growth[6] (Fig. 1; see supplementary information). This result is conservative because the minimum possible ratio value (representing maximum subtype structure) was used in the Congo analysis. Furthermore, although subtypes can be clearly identified in the distribution of pairwise distances for the global sequences (Fig. 1, inset), there is much less distinction between intra- and intersubtype comparisons for the Congo sequences. Hence, for any two randomly chosen Congo sequences, it is difficult to determine unambiguously whether they belong to the same or to different subtypes.

Our results indicate that the Congo and global phylogenies probably result from different epidemiological histories. As many Congo strains appear to be basal, we propose that each global subtype is the result of the chance exportation of some Congo strains to other geographical regions, thus producing an apparent starburst. Such founder effects have

© 2001 Macmillan Magazines Ltd

been proposed to explain the phylogenetically distinct subtypes B and E of HIV-1 group M (ref. 2). The observation that many Congo strains fall basal to the global subtypes also suggests that previous phylogenetic analysis has underestimated the number of lineages that pre-date 1957–60, and hence underestimated the minimum number of cross-species transmissions necessary to reconcile the OPV hypothesis with phylogenetic data.

In conclusion, the HIV-1 sequences from the Congo are evidence that the claim of the OPV theory[1] that it is "probably the only hypothesis of origin that can readily explain the starburst phenomenon" is incorrect. Our results give us no reason to doubt that the last common ancestor of HIV-1 group M was present in a human host.

Andrew Rambaut*, David L. Robertson*, Oliver G. Pybus*, Martine Peeters†, Edward C. Holmes*

**Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*
e-mail: edward.holmes@zoo.ox.ac.uk
†Laboratoire Retrovirus, IRD, BP 5045, 34032 Montpellier Cedex1, France

1. Hooper, E. *The River: A Journey to the Source of HIV and AIDS* (Penguin, London, 1999).
2. Korber, B. *et al. Science* **288,** 1789–1796 (2000).
3. Hahn, B., Shaw, G., de Cock, K. M. & Sharp, P. M. *Science* **287,** 607–614 (2000).
4. Vidal, N. *et al. J. Virol.* **74,** 10498–10507 (2000).
5. Los Alamos National Laboratory HIV Sequence Database, http://hiv-web.lanl.gov.
6. Pybus, O. G., Rambaut, A. & Harvey, P. H. *Genetics* **155,** 1429–1437 (2000).
7. Robertson, D. L. *et al. Science* **288,** 55–57 (2000).

Supplementary information is available on *Nature*'s website at www.nature.com or as paper copy from the London editorial office of *Nature*.

Chronobiology

Reversal of honeybee behavioural rhythms

Adult honeybees have sleep-like states[1,2] and, like human infants[3], bees develop their own endogenous circadian rhythms as they mature[4,5]. But whereas disruption of our sleep cycles and synchronized internal rhythms may adversely affect our physiology and performance[3], we show here that honeybees can revert to certain arrhythmic behaviours when necessary. To our knowledge, this chronobiological plasticity is the first example in any animal of a socially mediated reversal in activity rhythms.

Circadian rhythms in honeybees are an important component of the social behaviour development process that underlies the colony's division of labour. Larvae must be fed around the clock and are 'nursed' in the hive by young bees (5–15 days old) that work without any overt behavioural rhythms[6]. At about three weeks of age, however, a bee begins to forage outside the hive for pollen and nectar, an activity that calls for an internal circadian clock for timing visits to flowers and for sun-compass navigation[7].

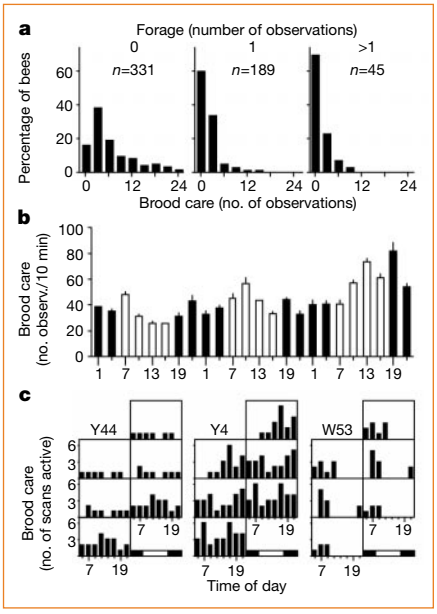


Figure 1 Reverted nurses care for brood with no diurnal rhythm. Brood care was observed under dim red light (invisible to bees[7]) every 3 h for three days. Observations of brood care[6]: six 10-min visual scans of individually tagged bees in the vicinity of the brood. Foraging observations were made as described[6]. **a**, Reorganization of division of labour in reversion colonies: frequency distributions of brood care differed significantly (chi-square test, $P<0.05$) for bees never observed foraging (0, left plot), observed foraging once (1, middle), or observed foraging more than once (>1, right). **b,** Colonial analysis. Mean (±s.e.) number of brood care events per observation period during the day (white bars) and night (black bars) ($n=6$ scans per observation). To test comprehensively for diurnal rhythms, we pooled the data into two half-day categories and compared the amount of brood-care activity between them; this analysis was repeated for eight different half-day combinations. No behavioural rhythms were detected ($P>0.05$, chi-square tests with Bonferonni correction). Results were similar for two other colonies (data not shown). Foragers and reverted nurses did not differ in age (29.9±0.2 days, $n=26$, and 29.8±0.4, $n=8$, respectively; $P=0.76$, unpaired *t*-test). **c,** Individual analyses. Number of scans with brood care (days double-plotted). Bars at the bottom right show the light–dark regime outside: black, night; white, day). Sixty-six reverted nurses were analysed individually. Y44: example of a bee active around the clock and showing no diurnal rhythm in brood care ($P>0.05$; statistical analyses as above); this behaviour was seen in 80.3% of reverted nurses. Y4: example of a bee active around the clock and with a weak diurnal rhythm in brood care ($P>0.05$); this behaviour was seen in 15.2% of reverted nurses. W53: one of only three bees (4.5%) showing clear diurnal rhythms ($P<0.05$).

Honeybees show great plasticity during their behavioural development, with their hive-to-field transition being accelerated, delayed, or even reversed in response to changing colony conditions[8]. We therefore investigated whether this plasticity extends to the bees' behavioural rhythms, focusing on the reversion from foraging to nursing as a particularly compelling challenge to the clock. This reversion occurs in response to a severe shortage of nurse bees and is associated with changes in exocrine, endocrine and neurochemical processes[8,9]. Do foragers induced to return to nursing also revert to an arrhythmic behavioural state?

We established three unrelated colonies, each composed initially of 2,000–2,500 foragers, their queen and young (sib) brood. Colonies composed only of foragers are known to induce behavioural reversion[8], and indeed the division of labour was reorganized in these colonies: many bees continued to forage, participating in little or no nursing behaviour; some foragers reverted to nursing and stopped foraging completely, or almost completely (Fig. 1a).

As in typical colonies with young nurses[6], brood care in our experimental colonies was performed around the clock, with no diurnal oscillations (Fig. 1b). The uninterrupted nursing occurred because individual bees had reverted to arrhythmic activity: analysis of individually tagged reverted nurses ($n=66$) revealed that brood care was performed by arrhythmic bees nursing day and night, rather than by rhythmic bees working in shifts (Fig. 1c). We found that reversion also affected the activity–rest cycle: 21 reverted bees (31.8%) cared for the brood in seven or more consecutive observations for 21 hours or longer; foragers, in contrast, rest daily for periods of seven hours or more[2].

The underlying cellular and molecular basis of this striking natural behavioural plasticity is unknown. There may be task-dependent changes in a central clock mechanism, uncoupling of nursing activity from the clock, or an effect resulting from nursing behaviour that overrides the clock output. Comparing these possibilities should help to clarify the nature of the cellular and molecular[4] bases of chronobiological plasticity.

Reverted nurses were able to rear the brood to maturity in all three colonies. Although we did not test other possible consequences of reversion, our findings may have wider implications, given the conservation of some molecular components of biological clocks[10] and of sleep regulation[11,12].

Guy Bloch*, Gene E. Robinson*†

**Department of Entomology, †Neuroscience Program, University of Illinois, 505 South Goodwin Avenue, Urbana, Illinois 61801, USA*
e-mail: guybloch@life.uiuc.edu

1. Kaiser, W. & Steiner-Kaiser, J. *Nature* **301,** 707–709 (1983).
2. Kaiser, W. *J. Comp. Physiol.* **163,** 565–584 (1988).
3. Turek, F.W. & Zee, P. C. *Regulation of Sleep and Circadian Rhythms* (Dekker, New York, 1999).
4. Toma, D. P., Bloch, G., Moore, D. & Robinson, G.E. *Proc. Natl Acad. Sci. USA* **97,** 6914–6919 (2000).
5. Stussi, T. & Harmelin, M. L. *C. R. Acad. Sci. Hebd. Seances. Acad. Sci. D* **262,** 2066–2069 (1966).
6. Moore, D., Angel, J. E., Cheeseman, I. M., Fahrbach, S. E. & Robinson, G. E. *Behav. Ecol. Sociobiol.* **43,** 147–160 (1998).
7. Von Firsch, K. *The Dance Language and Orientation of Bees* (Harvard University Press, Cambridge, 1967).
8. Huang, Z. Y. & Robinson, G. E. *Behav. Ecol. Sociobiol.* **39,** 147–158 (1996).
9. Schulz, D. J. & Robinson, G. E. *J. Comp. Physiol. A* **184,** 481–488 (1999).
10. Dunlap, J. C. *Cell* **96,** 271–290 (1999).
11. Shaw, P. J., Cirelli, C., Greenspan, R. J. & Tononi, G. *Science* **287,** 1834–1837 (2000).
12. Hendricks, J. C. *et al. Neuron* **25,** 129–138 (2000).

© 2001 Macmillan Magazines Ltd

Deep-mantle high-viscosity flow and thermochemical structure inferred from seismic and geodynamic data

Alessandro M. Forte* **& Jerry X. Mitrovica†**

* *Department of Earth Sciences, University of Western Ontario, Biology & Geology Building, London, Ontario, N6A 5B7 Canada*
† *Department of Physics, University of Toronto, 60 St George Street, Toronto, Ontario, M5S 1A7 Canada*

Surface geophysical data that are related to the process of thermal convection in the Earth's mantle provide constraints on the rheological properties and density structure of the mantle. We show that these convection-related data imply the existence of a region of very high effective viscosity near 2,000 km depth. This inference is obtained using a viscous-flow model based on recent high-resolution seismic models of three-dimensional structure in the mantle. The high-viscosity layer near 2,000 km depth results in a re-organization of flow from short to long horizontal length scales, which agrees with seismic tomographic observations of very long wavelength structures in the deep mantle. The high-viscosity region also strongly suppresses flow-induced deformation and convective mixing in the deep mantle. Here we predict compositional and thermal heterogeneity in this region, using viscous-flow calculations based on the new viscosity profile, together with independent mineral physics data. These maps are consistent with the anti-correlation of anomalies in seismic shear and bulk sound velocity in the deep mantle. The maps also show that mega-plumes in the lower mantle below the central Pacific and Africa are, despite the presence of compositional heterogeneity, buoyant and actively upwelling structures.

A quantitative model of thermal convective flow in Earth's mantle should satisfy a number of fundamental geophysical observations. In particular, such a model must accurately predict the principal surface data sets associated with the mantle convection process, namely: the global-scale free-air gravity anomalies which are accurately constrained by satellite data[1]; the observed tectonic-plate motions[2]; the dynamic topography on the core–mantle boundary (CMB), in particular the excess or dynamic ellipticity of the CMB inferred from space geodetic observations of the free-core nutation period[3,4]; and the dynamic topography at the surface which may be extracted from the observed topography by removing the contribution due to isostatic compensation of crustal thickness variations[5–8]. A successful mantle flow model must also be consistent with the three-dimensional structure in global seismic tomographic models, especially the observed dominance of very long wavelength heterogeneity in the bottom 1,000 km of the lower mantle[9–13]. This last observation has served as a basis for recent proposals of compositional stratification in the bottom half of the mantle[14] and thermochemical convection in the mantle[15–17]. For a recent survey of the arguments for deep-mantle compositional heterogeneity, from a mainly geochemical perspective, see ref. 18.

Thermal and compositional effects on seismic anomalies

The possible presence of significant chemical heterogeneity in the deep mantle may well bias our inferences of the temperature and density structure of convection from seismic tomographic models. This possibility is highlighted by recent seismic studies of the deep lower mantle, which reveal an unusually high ratio of shear velocity to compressional velocity anomalies[19] and a clear anti-correlation between perturbations of seismic shear velocity and acoustic (bulk-sound) velocity[20–23]. To understand the implications of this anti-correlation we first estimated the temperature and compositional derivatives of density and seismic wave velocities at 2,740 km depth, which corresponds to the top of the D″ layer in the PREM[24] seismic reference model (see Methods). We restricted our analysis of major-element compositional heterogeneity to the simplified chemical system $MgO–FeO–SiO_2$ and hence did not consider possible effects[25,26] due to variations in the less abundant oxides CaO and Al_2O_3, which each constitute 3–4% (by weight) of the bulk mantle composition[27]. From the estimated temperature derivatives (Table 1) it is clear that the seismic shear velocity (V_s) displays the greatest sensitivity to temperature perturbations (especially when anelastic effects are included), whereas the bulk-sound velocity (V_ϕ) is an order of magnitude less sensitive to the temperature perturbations.

Inferences of mantle viscosity and density perturbations

Given the high sensitivity of seismic shear velocity to temperature variations, we selected two recent three-dimensional mantle models[13,28] that provide lateral variations of shear velocity and used them to construct convective-flow models which satisfy the convection-related surface data described above. These tomographic models are derived from distinct, independent data sets, using different parameterizations of the heterogeneity, and using different inversion procedures. Thus we may view these two three-dimensional seismic models as representative of a wide spectrum of possible models, with one[13] (here referred to as the 'Grand' model) based on the inversion of travel-time data and the other[28] (here referred to as the 'Ek&Dz' model) based mainly on the inversion of full seismic waveforms. We assumed initially that the seismic anomalies in the two models are thermal in origin. To begin, we therefore used a velocity-to-density scaling coefficient (long-dashed line, Fig. 1a) derived by Karato[29] and subsequently modified on the

Table 1 Seismic temperature and compositional derivatives at 2,740 km depth

Elastic property	Temperature derivatives, $\partial/\partial T$ ($\times 10^5$ K^{-1})		Compositional derivatives	
	Anharmonic	Anelastic	$\partial/\partial X_{Fe}$	$\partial/\partial X_{Pv}$
$\ln\rho$	−1.0	–	+0.32	$+4.3\times10^{-3}$
$\ln V_s$	−4.7	−2.4	−0.22	+0.045
$\ln V_p$	−2.2	−0.9	−0.18	+0.047
$\ln V_\phi$	−0.7	0	−0.16	+0.048

For details on the construction of this table see the Methods section.

© 2001 Macmillan Magazines Ltd

basis of geodynamic constraints[5]. The density perturbations we obtained using this scaling are summarized in Fig. 1b, which displays the average amplitude of $\delta\rho/\rho$ as a function of depth. With the exception of the top 300 km, we note that the two seismic models yield rather different amplitudes for the density perturbations. We use these inferences of density heterogeneity to determine directly the buoyancy forces that drive the mantle convective flow.

The theory we use to calculate the mantle flow that is expected on the basis of the tomography-derived density anomalies requires a model of the rheological properties of the mantle, which we represent here in terms of a depth-dependent effective viscosity. In this flow theory, the buoyancy-induced flow in the mantle is coupled to the 13 principal surface tectonic plates[2], treated as rigid bodies, whose motions are determined by the underlying flow field[30]. The viscous-flow theory assumes an effectively linear, newtonian viscosity which is strongly suggested for the lower mantle by high-temperature creep experiments on perovskite-analogue minerals[31].

The required model of effective viscosity has been obtained using a nonlinear, iterative, Occam-style inversion[32–35] of the main convection-related geophysical data. Figure 2a shows results from inversions based on the two three-dimensional seismic models and Table 2 (rows indicated by asterisks) summarizes the fit of the corresponding flow models to the data. The viscosity profiles (Fig. 2a) are very similar, despite the non-negligible differences in the amplitudes of the density contrasts delivered by the two seismic models (Fig. 1b). The inferred viscosity variations display a pronounced low-viscosity channel in the upper mantle (between 100 and 300 km depth) and two viscosity maxima in the lower mantle, one near the top of the lower mantle and the other at 2,000 km depth. These lower-mantle viscosity maxima are not artefacts of the inversion procedure because the Occam approach explicitly penalizes radial roughness in the viscosity profiles. We carried out tests in which we removed either of the two viscosity peaks, or simply joined them to create a single high-viscosity region, and in each case the fit to the convection data (in particular the gravity anomalies) was significantly reduced.

An alternative demonstration of how the convection data constrains mantle viscosity at different depths is provided by a resolving power analysis. The resolution kernels shown in Fig. 2b clearly indicate that the viscosity maxima near 1,000 km and 2,000 km depth are well resolved, with little trade-off with viscosity structure elsewhere. The viscosity minima near 200 km and 1,400 km depth are similarly well constrained. The viscosity resolution in the top half of the lower mantle is mainly provided by the plate-motion data whereas in the bottom half of the mantle (for example, at 2,000 km depth) the resolution is strongly controlled by the gravity-anomaly data (Fig. 2b).

Independent verification of the viscosity profiles in Fig. 2a is provided by a comparison with radial profiles of seismic attenuation or Q factor[36]. We calculated the average viscosity in each of the five mantle layers used to parameterize a recent seismic Q model[36]. Assuming the frequency (ω) dependence[37], $Q \approx \omega^{\alpha}$, we translated the five-layer viscosity model into an equivalent Q model and obtained the best least-squares fit to the actual seismic Q model using $\alpha = 0.275$, which falls in the currently accepted range[37] $\alpha = 0.1$–0.3.

Next, we consider the possibility of further improving the fit to the convection data by inverting the data to derive an optimal $d\ln\rho/d\ln V_s$ scaling coefficient, rather than adopting the *a priori* scaling (long-dashed line, Fig. 1a) derived from mineral physics data. We carried out Occam inversions for $d\ln\rho/d\ln V_s$, and the two profiles we obtained are displayed in Fig. 1a. The fits to the convection data provided by the new inferences are summarized in Table 2 (rows indicated with daggers). The match we obtained with the excess CMB ellipticity (Table 2) is essentially perfect and, in the case of Ek&Dz, this is mainly achieved by adjusting the density heterogeneity inferred in the lowermost mantle using the $d\ln\rho/d\ln V_s$ scaling (solid line, Fig. 1a). We note that with the new $d\ln\rho/d\ln V_s$ profiles, differences in the average amplitudes of the density perturbations (Fig. 1c) are diminished relative to Fig. 1b.

Dynamics of lower-mantle flow and implications for mixing

To explore the dynamical impact of the high viscosity peak at 2,000 km depth (Fig. 2a) we constructed a reference two-layer viscosity model by separately averaging the viscosity variations in the upper and lower mantle. This averaging procedure was applied to the viscosity profile inferred using the heterogeneity in the Ek&Dz model and the resulting profile (long-dashed line, Fig. 2a) is characterized by a factor of 50 jump in viscosity at 670 km depth.

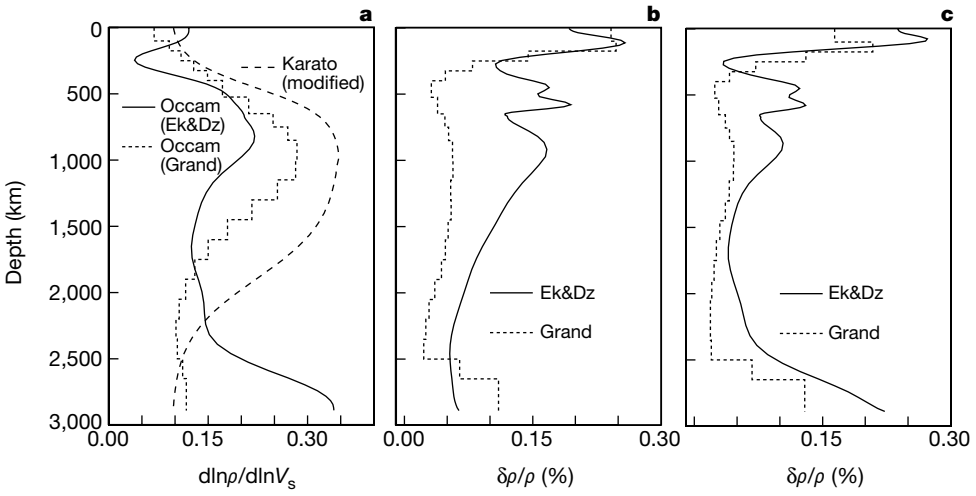


Figure 1 Geodynamic inferences of density structure in Earth's mantle. **a**, The long-dashed line represents the $d\ln\rho/d\ln V_s$ (velocity-to-density) scaling coefficient originally estimated by Karato[29] on the basis of mineral physics data that we subsequently modified on the basis of geodynamic data. The short-dashed and solid lines represent the $d\ln\rho/d\ln V_s$ scaling inferred for the three-dimensional seismic models of Grand[13] and Ekström and Dziewonski[28], respectively, after performing Occam inversions of the geodynamic data. (In these inversions the corresponding viscosity profiles in Fig. 2a were employed.) **b**, The root-mean-square (r.m.s.) amplitude of the density perturbations, as a function of depth, obtained from the three-dimensional seismic models of Grand[13] (dashed line) and Ekström and Dziewonski[28] (solid line). In both cases the $d\ln\rho/d\ln V_s$ (modified) scaling inferred by Karato[29] (long-dashed line in **a**) is employed. **c**, The r.m.s. amplitude of the density perturbations derived from the three-dimensional seismic models of Grand[13] (dashed line) and Ekström and Dziewonski[28] (solid line) on the basis of the respective Occam-inferred $d\ln\rho/d\ln V_s$ scalings in **a**.

© 2001 Macmillan Magazines Ltd

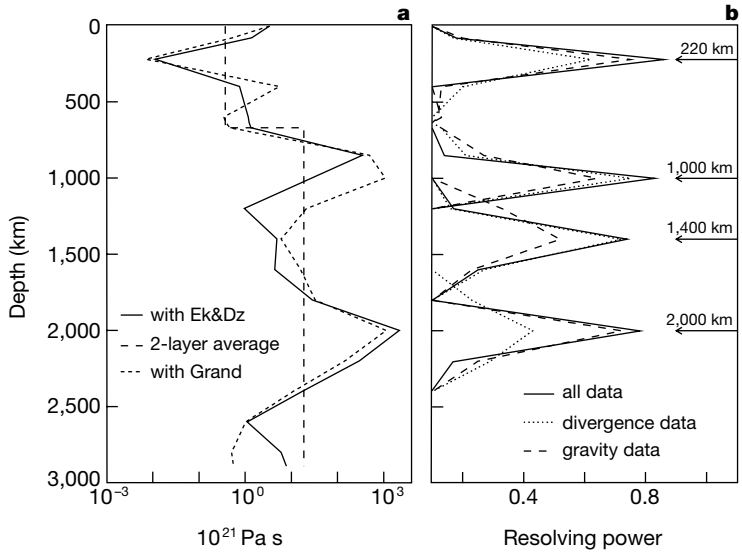


Figure 2 Depth-dependent effective viscosity of Earth's mantle. **a**, The short-dashed and solid lines represent the viscosity profiles inferred in Occam inversions of the geodynamic data (listed in Table 2) based on the seismic models of Grand[13] and Ekström and Dziewonski[28], respectively. In these Occam inversions the $d\ln\rho/d\ln V_s$ (modified) scaling of Karato[29] was employed (long-dashed curve in Fig. 1a). The long-dashed curve is a two-layer viscosity profile obtained by separately averaging the upper-mantle and lower-mantle viscosity variations in the profile (solid line) inferred on the basis of the Ekström and Dziewonski seismic model. **b**, The viscosity resolving power of the geodynamic data at four depths indicated by the horizontal arrows (for example, at 220 km). These curves, called 'resolution kernels', show the vertical averages of the mantle viscosity that are inherent in the viscosity inferences at these four depths. Perfect resolution would correspond to kernels with peak values of 1 and zero width. The solid lines show the resolution obtained with the full set of geodynamic data (gravity anomalies, plate divergence, and CMB ellipticity). The dotted lines show the resolution obtained when we ignore the gravity data and the dashed lines show the resolution in the absence of plate-divergence data.

The lower-mantle flow field predicted on the basis of this two-layer viscosity is depicted in Fig. 3a, c at depths of 1,300 and 1,800 km, respectively. Clearly, in this flow calculation the pattern of both the horizontal and vertical flow fields is unchanged across this depth range.

In contrast to this simple behaviour, consider Fig. 3b, d, where we plot flow predicted on the basis of the viscosity profile (solid line, Fig. 2a) derived by inversion of the convection data. We now observe a remarkable change in the flow regime between 1,300 and 1,800 km depth. At 1,300 km depth (Fig. 3b), the flow is organized on relatively short horizontal length scales, with clearly defined upwellings below the Pacific, especially below Hawaii and French Polynesia in the central Pacific. These upwelling centres are being fed by converging horizontal flow which, in the case of the Hawaiian plume, seems to originate from as far north as the Aleutian subduction zone. This observation seems to provide a natural explanation for the observed signature of recycled oceanic crust in ocean-island volcanic rocks[38]. In contrast, at 1,800 km depth (Fig. 3d) the individual upwellings are now absent and both the vertical and horizontal flow are dominated by very long horizontal wavelengths. Although the viscosity increases rapidly towards its maximum near 2,000 km depth, its control on the horizontal organization of flow extends to shallower depths and is already discernible at 1,800 km depth. The flow character at this depth is closely reminiscent of the oscillatory 'doming regime' observed in recent laboratory convection experiments[17].

A useful measure of the degree of mixing induced in laminar fluid flows is provided by the time-dependent separation or 'stretching' of originally adjacent particles in the fluid[39,40]. To quantify this stretching we determined the direction and magnitude of the principal (that is, maximum) strain rate at a large number of points distributed throughout the mantle, employing the strain-rate tensor $d\epsilon_{ij}/dt$ associated with the predicted flow fields. The largest of the principal stretching rates predicted at any given depth is shown in Fig. 4, where we immediately note a strong minimum at 2,000 km depth. Another prominent minimum is predicted at the top of the lower mantle, near 1,000 km depth. These zones of very low deformation rate, and hence reduced mixing, reflect the two high-viscosity regions in the inverted models; they are absent in the flow predicted using the simple two-layer viscosity profile (dashed line, Fig. 4).

Tomography-based studies of convection-related surface data have long argued for an increase of 1–2 orders of magnitude in the average viscosity between the upper and lower mantle[33,34,41–43]. These variations affect both the mixing properties[44–46] and the planform of mantle convective flow[47]. Recent investigations[46] suggest that viscosity stratification alone may not be sufficient to explain the trace-element geochemical constraints on mantle

Table 2 Fits to global convection data

Three-dimensional mantle model	Free-air gravity, $l = 2$–20	Plate divergence, $l = 1$–32	Excess CMB ellipticity	r.m.s. topography $l = 1$–20
Grand*	31% (54%)	60%	0.50 km	0.91 km
Grand†	32% (56%)	59%	0.52 km	0.71 km
Ek&Dz*	37% (77%)	57%	–0.16 km	1.32 km
Ek&Dz†	43% (71%)	69%	0.52 km	0.95 km
Ek&Dz‡	9% (–65%)	68%	0.76 km	1.04 km

All fits, with the exception of the surface and CMB topography, are expressed in terms of variance reduction. The observed free-air gravity anomalies are calculated using the (non-hydrostatic) geopotential derived from satellite data[1]. The plate-divergence data are represented here in terms of the horizontal divergence of the plate velocities given by the NUVEL-1 model[2]. The most recent inference[4] of excess or dynamic CMB ellipticity suggests a value of about 0.4 km, rather than 0.5 km as determined in earlier studies[3]. (This 20% difference is well within the uncertainties associated with the three-dimensional seismic models used in these flow predictions.) The dynamic surface topography is not employed in the Occam inversions for the mantle viscosity profile. The root-mean-square (r.m.s.) amplitude of the dynamic topography is used as an independent *a posteriori* check of the plausibility of the convective-flow models. The most recent inference[5] of the r.m.s. amplitude of the dynamic surface topography is 0.9 km. The numbers in parentheses are the variance reductions to the equivalent non-hydrostatic geoid anomalies derived from satellite data[1]. The wavelength-dependent amplitude spectra of the geoid and gravity anomalies are quite different; the latter has a relatively 'flat' spectrum, which explains why the fits to the anomalous gravity and geoid fields differ.

* The density perturbations in the convective-flow models are derived from the shear-velocity anomalies assuming a thermal origin (using the $d\ln\rho/d\ln V_s$ 'Karato' profile in Fig. 1a).

† These density perturbations are instead derived on the basis of the associated $d\ln\rho/d\ln V_s$ profiles (Fig. 1a) inferred in Occam inversions of the convection data.

‡ The two-layer viscosity model (Fig. 2a) along with the Occam-inferred $d\ln\rho/d\ln V_s$ profile (solid line, Fig. 1a) are used in this viscous flow calculation.

© 2001 Macmillan Magazines Ltd

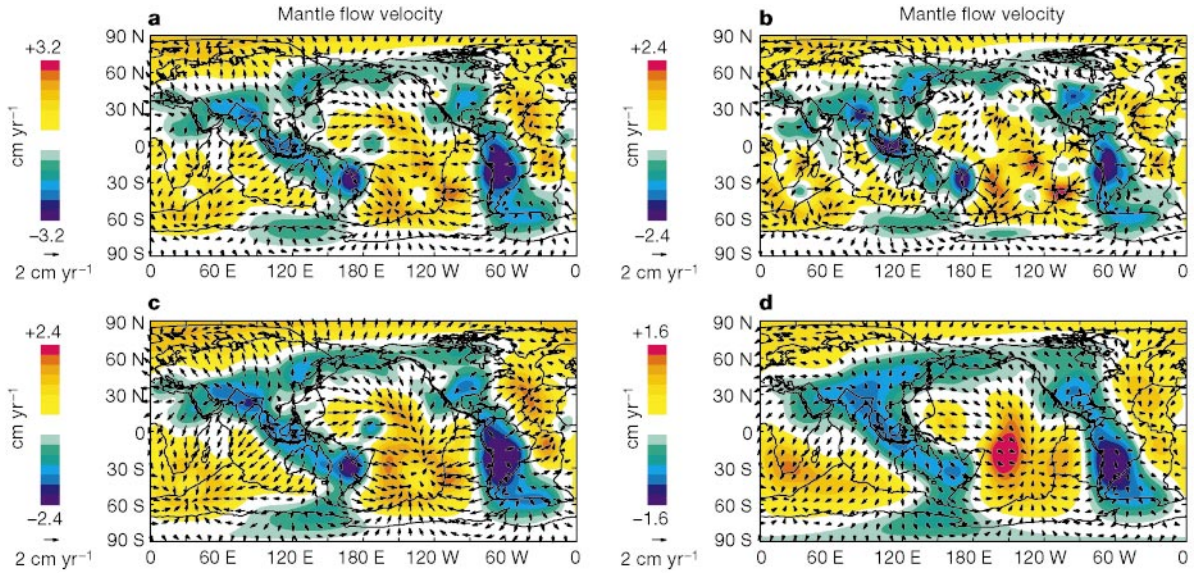


Figure 3 Predicted convective flow in Earth's lower mantle. In all maps the colour contours represent the speed of vertical flow (see the colour scale bars in cm yr^{-1}) with the blue colours representing regions of downward flow and the orange colours representing regions of upward flow. The black arrows represent the speed and direction of the horizontal component of the flow (see the horizontal velocity scale arrow). All maps depict the convective flow predicted on the basis of the mantle density anomalies derived from the three-dimensional seismic model of Ekström and Dziewonski[28] (using the $d\ln\rho/d\ln V_s$ scaling represented by the solid line in Fig. 1a). The maps in **a** and **c** represent the flow field predicted on the basis of the two-layer viscosity profile (long-dashed curve in Fig. 2a). The maps in **b** and **d** represent the flow field predicted on the basis of the Occam-inferred viscosity profile (solid curve in Fig. 2a). **a**, **b**, Depth of 1,300 km. **c**, **d**, Depth of 1,800 km.

mixing. However, the viscosity profiles employed in these numerical convection studies were not constrained to fit the full suite of surface geophysical constraints (in particular the global gravity anomalies and the pattern of present-day tectonic-plate velocities).

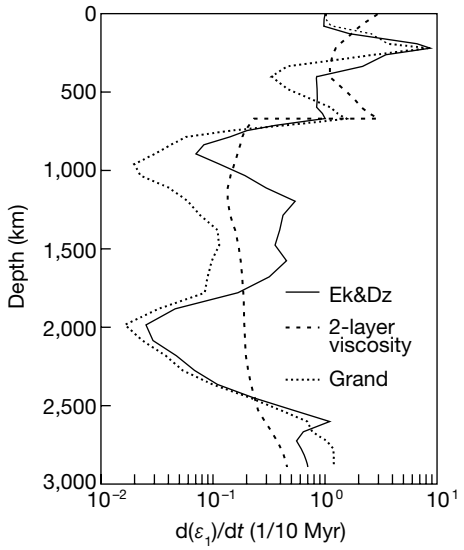


Figure 4 Flow-induced deformation or 'stretching' in the mantle. The maximum value, at any depth, of the principal rate of strain (defined as the largest positive eigenvalue of the strain-rate tensor $d\epsilon_{ij}/dt$) is plotted as a function of depth. The dashed line represents the extension rate associated with the flow field (Fig. 3a, c) predicted with the two-layer viscosity model (Fig. 2a) and the density contrasts (Fig. 1c) derived from the three-dimensional seismic model of Ekström and Dziewonski[28]. The solid line represents the extension rate associated with the flow field (Fig. 3b, d) predicted with the Occam-inferred viscosity (Fig. 2a) and the density contrasts (Fig. 1c) derived from the Ekström and Dziewonski seismic model. The dotted line is the flow-induced extension rate obtained on the basis of the density contrasts (Fig. 1c) derived from the three-dimensional seismic model of Grand[13] and the corresponding Occam-inferred viscosity (Fig. 2a).

Indeed, an idealized two-layer representation of the mantle viscosity, commonly employed in the numerical simulations, provides a poor fit to the convection data (see Table 2) and is unable to produce complex depth-dependent mixing behaviour (Fig. 4).

The flow calculations illustrated in Fig. 3 show that the transition from a domed pattern of flow in the deep lower mantle to a pattern of individual upwellings in the mid-mantle, long suggested by tomographic models[9–13], may largely be explained in terms of viscosity stratification. We should not conclude, however, that viscosity alone is responsible for maintaining this doming and that compositional heterogeneity is thus unnecessary or absent in the deep lower mantle, especially in view of the observed anti-correlation of seismic shear and bulk-sound velocity anomalies in this region[20–23]. We next consider the implications of this anti-correlation.

Thermochemical heterogeneity in the deep lower mantle

Recent tomographic studies which yield a joint description of seismic shear and bulk-sound velocity anomalies[20,22] provide a new means for determining the relative contributions of thermal and compositional heterogeneity in the deep mantle. Using Table 1, we have derived the following expressions (see Methods) relating specific linear combinations of shear and bulk-sound velocity anomalies to effective measures of compositional and thermal heterogeneity at 2,740 km depth:

$$-2.3\delta\ln V_s + 23.0\delta\ln V_\phi = \delta X_{Pv} - 3.2\delta X_{Fe} \equiv \delta X_{eff} \quad (1)$$

$$(-15.5\delta\ln V_s + 14.5\delta\ln V_\phi)\times 10^3 = \delta T + 1086\delta X_{Fe} \equiv \delta T_{eff} \quad (2)$$

The effective compositional heterogeneity δX_{eff} reflects lateral variations in the content of silica (via changes in the molar fraction of perovskite, δX_{Pv}) and iron (δX_{Fe}). The effective thermal heterogeneity δT_{eff} is a combined representation of temperature perturbations and lateral variations in iron content; however, we demonstrate below (see also Methods) that δT_{eff} provides an accurate approximation for δT.

Estimates of the effective compositional and thermal anomalies at 2,740 km depth, derived from two independent models of

© 2001 Macmillan Magazines Ltd

combined shear and bulk sound velocity anomalies[20,22], are in excellent agreement (Fig. 5). These estimates clearly show that δX_{eff} is strongly correlated to bulk-sound velocity anomalies while the shear velocity anomalies are strongly anti-correlated to δT_{eff}. We have tested the robustness of these results by considering a third, recently derived seismic model of lower mantle structure[48], and we obtained maps of δX_{eff} and δT_{eff} (see Supplementary Information) which also agree well with those shown here.

Our estimates of δX_{eff} and δT_{eff} were based on results from seismology and mineral physics. To extend this analysis to estimate lateral variations of iron or perovskite content requires an independent constraint on density perturbations. Using Table 1, we can write (see Methods):

$$\delta X_{\mathrm{Fe}} = 2.9(\delta \ln \rho + 10^{-5}\delta T_{\mathrm{eff}}) - 0.013\delta X_{\mathrm{eff}} \qquad (3)$$

We will use the $\delta\ln\rho$ field determined from geodynamic modelling (Fig. 1c). This approach assumes that the estimated amplitude of $\delta\ln\rho$ variations is independent of the initial $\delta\ln V_s$ model adopted in the geodynamic modelling and that a radially dependent scaling from shear velocity to density is reasonably valid. The former is demonstrated in Fig. 1c, while the latter is supported by the fits to the geodynamic observables[5].

The maps of iron heterogeneity (Fig. 6c–f) all show peak anomalies of about $\delta X_{\mathrm{Fe}} = \pm 0.01$, equivalent to about 10% variation relative to the bulk average value for X_{Fe}. Following equation (2), this level of iron heterogeneity suggests an error of the order of 10 K in using δT_{eff} as a proxy for δT. Accordingly, we have estimated the effective temperature derivative of shear velocity, $(\mathrm{d}\ln V_s/\mathrm{d}T)_{\mathrm{eff}}$, which provides an optimal fit (in the least-squares sense) between the δT_{eff} and $\delta\ln V_s$ fields in Fig. 5. Our estimate of $-5.2\times10^{-5}\,\mathrm{K}^{-1}$ is smaller than the purely thermal value of $-7.1\times10^{-5}\,\mathrm{K}^{-1}$ (Table 1) because of the 'masking' effect of compositional heterogeneity.

Lateral variations in chemical composition represented by δX_{Fe}

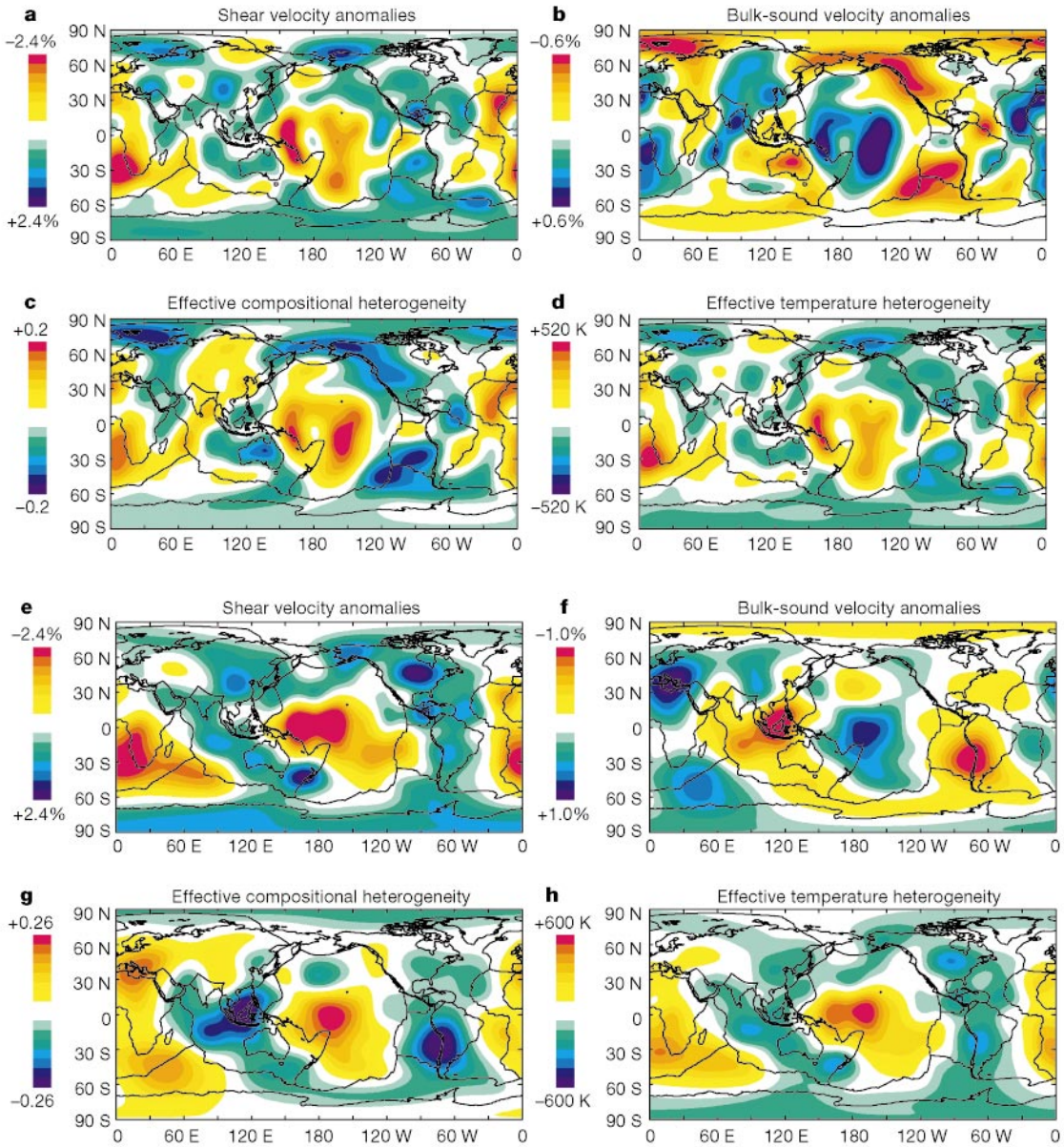


Figure 5 Compositional and thermal heterogeneity at 2,740 km depth. **a**, Relative shear velocity anomalies $\delta V_s/V_s$ obtained from the Su and Dziewonski (Su&Dz) tomography model[20]. **b**, Relative bulk-sound velocity anomalies $\delta V_\phi/V_\phi$ from Su&Dz. **c**, The effective compositional heterogeneity δX_{eff} derived from the Su&Dz model using equation (1). **d**, The effective temperature perturbations δT_{eff} derived from the Su&Dz model using equation (2). The maps in **e**–**h** are analogous to the maps in **a**–**d**, respectively, except that in **e**–**h** we used the joint description of shear and bulk-sound velocity anomalies given by the tomography model of Masters *et al.*[22]

© 2001 Macmillan Magazines Ltd

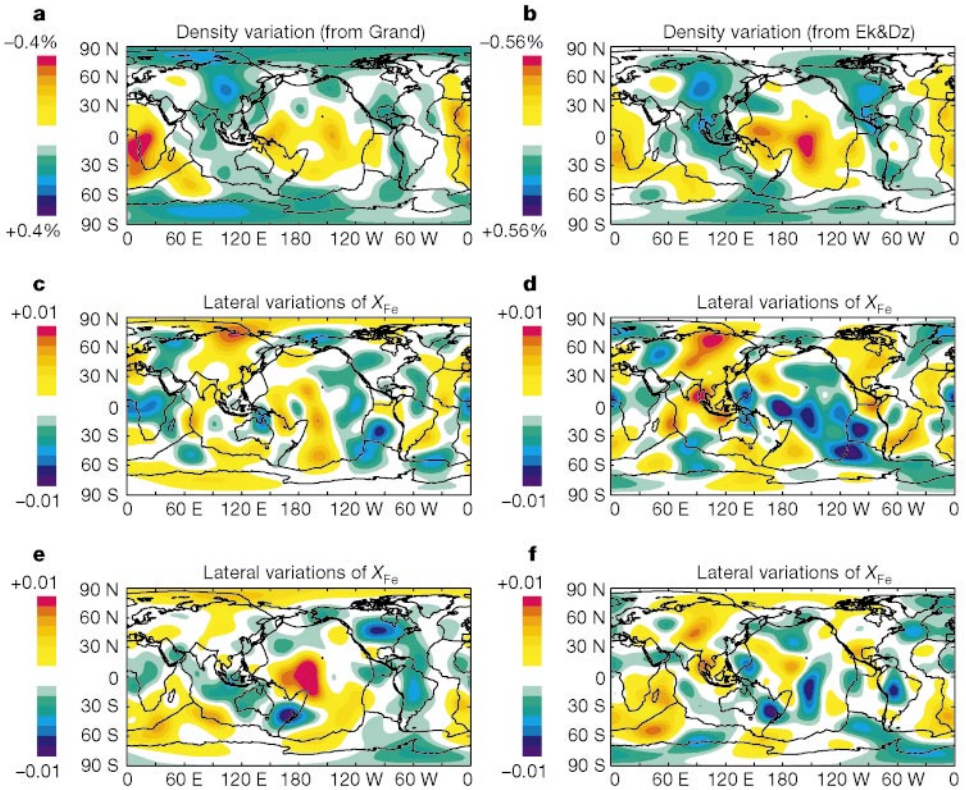


Figure 6 Lateral variations in iron content at 2,740 km depth. The inferences of iron heterogeneity δX_{Fe} shown here are obtained on the basis of equation (3), using δX_{eff} and δT_{eff} in Fig. 5, and using the geodynamic inferences of density heterogeneity (Fig. 1c). The density heterogeneity maps **a** and **b**, were obtained from the tomography models of Grand[13] and Ekström and Dziewonski[28], respectively. In maps **c** and **e** are shown δX_{Fe} derived using the tomography models of Su and Dziewonski[20] (Su&Dz) and Masters *et al.*[22] (Scripps), respectively. In both cases the inferences of δX_{Fe} are based on the density heterogeneity to map **a**. In maps **d** and **f** are shown δX_{Fe} again derived using the Su&Dz and Scripps models, respectively, except that the density heterogeneity in map **b** is used.

(Fig. 6) produce associated density anomalies $(\delta \ln\rho)_{chem}$ that may oppose the pure thermal buoyancy $(\delta \ln\rho)_{th} = -\alpha\delta T$, where α is the thermal expansivity. The magnitude of this effect, quantified by the buoyancy ratio $R_\rho = -(\delta \ln\rho)_{chem}/(\delta \ln\rho)_{th}$, has been shown to exert fundamental control on the dynamics of convection in the lower mantle[15–17]. Using Figs 5 and 6 we calculated $(\delta \ln\rho)_{chem}$ and $(\delta \ln\rho)_{th}$ and then determined the buoyancy ratio R_ρ which provides an optimal fit between these two fields. For δX_{Fe} extracted from density anomalies derived from the Grand (Fig. 6a) and Ek&Dz (Fig. 6b) models we find approximate values for R_ρ of 0.3 and 0.1, respectively. These values are significantly smaller than the $R_\rho \approx 1$ which has been assumed in some numerical convection simulations[15,16], but they fall well within the parameter space associated with a distinct doming regime[17] for mantle convection.

In view of the large-amplitude lateral temperature contrasts we have inferred (Fig. 5d, h), a legitimate concern is the extent to which the associated lateral variations in viscosity may affect our interpretations based on a purely depth-dependent effective viscosity. We constructed a tomography-based mantle flow model which explicitly incorporates lateral viscosity variations at all depths, using a previously developed theoretical formulation[30]. The lateral viscosity heterogeneity we employ is derived using a homologous temperature scaling[31], yielding large-scale lateral variations which range over three orders of magnitude (see Supplementary Information). These lateral variations are comparable to the magnitude of the geodynamically inferred radial variations in viscosity (Fig. 2a). We find that the relative mean-square amplitude of the differences between predictions (for example, geoid anomalies, CMB topography) obtained with and without these lateral viscosity variations (see Supplementary Information) are not more than 10%, which is smaller than the current misfit to the convection-related data (Table 1). Thus the error which may be incurred by neglecting large-scale lateral viscosity heterogeneity is no larger than other uncertainties (for example, in the seismic tomographic models) inherent in the current tomography-based viscous-flow models.

We conclude from Fig. 6a, b that the combined thermal and chemical anomalies in the deep lower mantle are such that the hot mega-plumes (Fig. 5d, h) below the central Pacific, and below Africa, are positively buoyant and hence actively ascending. Our analysis indicates that their buoyancy is partially reduced by the presence of compositional heterogeneity (Fig. 6c–f), but the temperature effect is dominant. This result is in agreement with other recent investigations of the dynamics of these deep mantle plumes. For example, the uplift rates for southern Africa, estimated from geological data, suggest[49] that the deep-mantle African mega-plume must be positively buoyant in order to match the estimated time-dependent uplift during the Cenozoic era. The global geophysical constraints on the dynamics of these deep mantle mega-plumes thus seem to suggest consistently that these structures are buoyant and that their dynamics (Fig. 3) are strongly controlled by the high-viscosity region that we infer near 2,000 km depth. □

Methods

Construction of Table 1

The (anharmonic or purely elastic) temperature (T) derivatives were estimated on the basis of the equation-of-state modelling by Jackson[50], in particular the results (see Fig. 9c, d in ref. 50) concerning the temperature-dependence of density ρ and seismic parameter $\phi = V_\phi^2$, where V_ϕ is the bulk-sound velocity. For example, at pressure $P = 127$ GPa (2,740 km depth), a variation in potential temperature from $T_0 = 1{,}600$ K to $T_0 = 1{,}850$ K results in a 0.44% change in density. Because $\partial/\partial T \approx (T_0/T)\partial/\partial T_0$, as shown by Stacey[51], we estimated a thermal expansion coefficient $\alpha = 1 \times 10^{-5}\ \mathrm{K}^{-1}$. From the variation in ϕ determined by Jackson[50], we similarly obtained the logarithmic T-derivative of the bulk-sound velocity $\partial \ln V_\phi/\partial T$. To determine the T-derivative of seismic shear (V_s)

© 2001 Macmillan Magazines Ltd

and compressional velocity (V_p) we also required an estimate of the T-dependence of the elastic shear modulus μ. Using Stacey's[51] derivation for $\Gamma = -(\alpha\mu)^{-1}(\partial\mu/\partial T)_p$ (see equations (70)–(71) in ref. 51) we determined the T-derivatives of V_s and V_p.

Following Karato[29], we estimated the anelastic contribution to the T-derivatives by assuming a frequency-independent seismic attenuation, or Q factor: $(\partial \ln V_s/\partial T)_{\text{anelastic}} = -(\pi Q_\mu T)^{-1}(H/RT)$, where H is the (P-dependent) activation enthalpy. Using Weertman's rule, the activation enthalpy is related to melting temperature T_m using $H = gRT_m$, where R is the gas constant and g is estimated to be approximately 30, a value appropriate for magnesium silicates such as olivine[52]. For T_m we employed a recent laboratory estimate of the deep-mantle pyrolite solidus[53]. We also used $Q_\mu = 312$, according to PREM[24], and assumed that all attenuation occurs in shear deformation (that is, the bulk attenuation is assumed to be zero).

The compositional derivatives represent the changes in density and seismic velocity due to changes in the molar fraction of iron, $X_{Fe} = [FeO]/([FeO] + [MgO])$, and the molar fraction of perovskite, $X_{Pv} = [Pv]/([Pv] + [Mw])$, where Mw denotes magnesio-wüstite. The molar fraction of perovskite is equivalent to the silica ratio: $X_{Pv} = [SiO_2]/([FeO]+[MgO])$. Variations in density and bulk-sound velocity caused by variations in iron content are (to a very good approximation) independent of pressure, and hence depth (see Fig. 9e, f in ref. 50). We thus calculated the derivatives $\partial/\partial X_{Fe}$ from standard temperature and pressure laboratory data (for example, Table 1 in ref. 50). The derivatives with respect to perovskite or silica content, $\partial/\partial X_{Pv}$, do vary with depth and we estimated the derivatives for ρ and V_ϕ directly from results in Fig. 9e, f in ref. 50. Derivatives of the shear modulus with respect to X_{Pv} are derived on the basis of the expression $\mu = (\mu/\kappa)_0[\kappa - \kappa'_\infty P]$ obtained by Stacey[54]. This relation allows us to express the compositional derivatives of shear modulus in terms of the derivatives of the bulk modulus, thereby yielding values for $\partial \ln V_s/\partial X_{Pv}$ and $\partial \ln V_p/\partial X_{Pv}$.

Estimating thermal and compositional variations

We begin with the following general expressions, based on the parameters in Table 1:

$$\delta \ln V_s = \frac{\partial \ln V_s}{\partial T}\delta T + \frac{\partial \ln V_s}{\partial X_{Pv}}\delta X_{Pv} + \frac{\partial \ln V_s}{\partial X_{Fe}}\delta X_{Fe}$$

$$\delta \ln V_\phi = \frac{\partial \ln V_\phi}{\partial T}\delta T + \frac{\partial \ln V_\phi}{\partial X_{Pv}}\delta X_{Pv} + \frac{\partial \ln V_\phi}{\partial X_{Fe}}\delta X_{Fe}$$

$$\delta \ln \rho = \frac{\partial \ln \rho}{\partial T}\delta T + \frac{\partial \ln \rho}{\partial X_{Pv}}\delta X_{Pv} + \frac{\partial \ln \rho}{\partial X_{Fe}}\delta X_{Fe}$$

where $\delta \ln V_s$, $\delta \ln V_\phi$ and $\delta \ln \rho$ are relative perturbations to shear velocity, bulk sound velocity, and density, respectively. We rewrite this system of equations in the following symbolic form:

$$\begin{pmatrix} \delta \ln V_s \\ \delta \ln V_\phi \\ \delta \ln \rho \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix} \begin{pmatrix} \delta T \\ \delta X_{Pv} \\ \delta X_{Fe} \end{pmatrix} \quad (4)$$

As joint seismic models for V_s and V_ϕ presently exist, we initially focus on the first two rows of equation (4). Eliminating δT in these rows yields:

$$\delta X_{\text{eff}} = \frac{c_{21}}{c_{12}c_{21} - c_{11}c_{22}}\delta \ln V_s - \frac{c_{11}}{c_{12}c_{21} - c_{11}c_{22}}\delta \ln V_\phi \quad (5)$$

where the 'effective' lateral variations in composition, δX_{eff}, are defined as:

$$\delta X_{\text{eff}} \equiv \delta X_{Pv} + \frac{c_{21}c_{13} - c_{11}c_{23}}{c_{12}c_{21} - c_{11}c_{22}}\delta X_{Fe} \quad (6)$$

We can furthermore show that these two rows can be expressed as:

$$\begin{pmatrix} \delta \ln V_s \\ \delta \ln V_\phi \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \begin{pmatrix} \delta T_{\text{eff}} \\ \delta X_{\text{eff}} \end{pmatrix}, \quad (7)$$

where we define the 'effective' temperature variation as:

$$\delta T_{\text{eff}} \equiv \delta T + \frac{c_{12}c_{23} - c_{13}c_{22}}{c_{12}c_{21} - c_{11}c_{22}}\delta X_{Fe}. \quad (8)$$

Inverting the matrix in equation (7) yields:

$$\begin{pmatrix} \delta T_{\text{eff}} \\ \delta X_{\text{eff}} \end{pmatrix} = \frac{1}{c_{12}c_{21} - c_{11}c_{22}} \begin{pmatrix} -c_{22} & c_{12} \\ c_{21} & -c_{11} \end{pmatrix} \begin{pmatrix} \delta \ln V_s \\ \delta \ln V_\phi \end{pmatrix}. \quad (9)$$

For a depth of 2,740 km we find, using Table 1:

$$\delta X_{\text{eff}} \equiv \delta X_{Pv} - 3.2\delta X_{Fe} = -2.3\delta \ln V_s + 23.0\delta \ln V_\phi \quad (10)$$

and

$$\delta T_{\text{eff}} \equiv \delta T + 1086\delta X_{Fe} = (-15.5\delta \ln V_s + 14.5\delta \ln V_\phi) \times 10^3 \quad (11)$$

The term $1086\delta X_{Fe}$ in equation (11) is the error introduced if we were to adopt δT_{eff} as an estimate for δT. The current best estimate[50] for the average mole fraction of iron in the lower mantle is $X_{Fe} = 0.11$, so we see from equation (11) that even if we assumed extreme lateral variations in iron content which approach 100% (that is, $\delta X_{Fe} = \pm 0.1$) the resulting contribution to δT_{eff} would be about 100 K, which is significantly smaller than the overall amplitude of δT_{eff} (Fig. 5d, h). The analysis described in the text (see Fig. 6) suggests peak variations of $\delta X_{Fe} \simeq 0.01$, and therefore the error in assuming $\delta T_{\text{eff}} \approx \delta T$ is only about 10 K. Variations in temperature and 'effective' composition in Fig. 5 are obtained by adopting the combination of seismically determined models for $\delta \ln V_s$ and $\delta \ln V_\phi$ specified in equations (10) and (11).

To separate the contributions to the effective variation δX_{eff} from δX_{Pv} and δX_{Fe} we incorporate density variations $\delta \ln \rho$ into the analysis. Using the last row of equation (4), together with the transformations (6) and (8), one can show that:

$$\delta \ln \rho = c_{31}\delta T_{\text{eff}} + c_{32}\delta X_{\text{eff}} - \frac{\Lambda}{c_{12}c_{21} - c_{11}c_{22}}\delta X_{Fe} \quad (12)$$

where Λ is the determinant of the matrix in equation (4). Given δT_{eff} and δX_{eff} from equation (9), we solve equation (12) for δX_{Fe}. Using the values in Table 1, the form of equation (12) appropriate for a layer at depth 2,740 km is:

$$\delta \ln \rho = -10^{-5}\delta T_{\text{eff}} + 0.0043\delta X_{\text{eff}} + 0.34\delta X_{Fe} \quad (13)$$

We have applied equation (13) to generate maps of δX_{Fe}, shown in Fig. 6.

The estimation of δT, δX_{Fe} and δX_{Pv} at a depth of 2,740 km is possible, in principle, because the matrix appearing in equation (4) is well conditioned. In practice, this exercise requires three independent inputs. The first two inputs are provided by the joint seismic models for $\delta \ln V_s$ and $\delta \ln V_\phi$. The final input, $\delta \ln \rho$, is determined on the basis of our geodynamic modelling (Fig. 1). Although we estimate the field $\delta \ln \rho$ by scaling $\delta \ln V_s$ to yield a fit to the geodynamic observables, we have shown that its amplitude is relatively insensitive to the starting model for $\delta \ln V_s$ (Fig. 1c). Thus, the geodynamic data provide the dominant constraint on the amplitude of $\delta \ln \rho$.

Received 10 April 2000; accepted 16 March 2001.

1. Marsh, J. G. *et al.* The GEM-T2 gravitational model. *J. Geophys. Res.* **95,** 22043–22071 (1990).
2. De Mets, C. R., Gordon, R. G., Argus, D. F. & Stein, S. Current plate motions. *Geophys. J. Int.* **101,** 425–478 (1990).
3. Gwinn, C. R., Herring, T. A. & Shapiro, I. I. Geodesy by radio interferometry: studies of the forced nutations of the Earth. 2. Interpretation. *J. Geophys. Res.* **91,** 4755–4765 (1986).
4. Mathews, P., Buffet, B. A. & Herring, T. A. What do nutations tell us about the Earth's interior? *Eos Trans. AGU* **80,** (Fall Meeting suppl.) 19 (1999).
5. Forte, A. M. & Woodward, R. L. Global 3D mantle structure and vertical mass and heat transfer across the mantle from joint inversions of seismic and geodynamic data. *J. Geophys. Res.* **102,** 17981–17994 (1997).
6. Mooney, W. D., Laske, G. & Masters, T. G. CRUST 5.1: A global crustal model at 5°×5°. *J. Geophys. Res.* **103,** 727–747 (1998).
7. Panasyuk, S. V. The effect of compressibility, phase transformations, and assumed density structure on mantle viscosity inferred from Earth's gravity field. PhD dissertation, Massachusetts Institute of Technology (1998).
8. Forte, A. M. & Perry, H. K. C. Geodynamic evidence for a chemically depleted continental tectosphere. *Science* **290,** 1940–1944 (2000).
9. Woodward, R. L. *et al.* in *Evolution of the Earth and Planets. Geophys. Monogr. Ser.* (eds Takahashi, E., Jeanloz, R. & Rubie, D.) Vol. 74, 89–109 (AGU, Washington, DC, 1993).
10. Masters, T. G. *et al.* A shear-velocity model of the mantle. *Phil. Trans. R. Soc. Lond. A* **354,** 1385–1411 (1996).
11. Li, X.-D. & Romanowicz, B. Global mantle shear-velocity model developed using nonlinear asymptotic coupling theory. *J. Geophys. Res.* **101,** 22245–22272 (1996).
12. van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. Evidence for deep mantle circulation from global tomography. *Nature* **386,** 578–584 (1997).
13. Grand, S. P., van der Hilst, R. D. & Widiyantoro, S. Global seismic tomography: A snapshot of convection in the Earth. *Geol. Soc. Am. Today* **7,** 1–7 (1997).
14. van der Hilst, R. D. & Kárason, H. Compositional heterogeneity in the bottom 1000 km of Earth's mantle: Towards a hybrid convection model. *Science* **283,** 1885–1888 (1999).
15. Tackley, P. J. in *The Core-Mantle Boundary Region.* Geodyn. Ser. (ed. Gurnis, M. *et al.*) Vol. 28, 231–253 (AGU, Washington, DC, 1998).
16. Kellogg, L. H., Hager, B. H. & van der Hilst, R. D. Compositional stratification in the deep mantle. *Science* **283,** 1881–1884 (1999).
17. Davaille, A. Simultaneous generation of hotspots and superswells by convection in a heterogeneous planetary mantle. *Nature* **402,** 756–760 (1999).
18. Albarède, F. & van der Hilst, R. D. New mantle convection model may reconcile conflicting evidence. *Eos Trans. AGU* **80,** 535–539 (1999).
19. Roberston, G. S. & Woodhouse, J. H. Constraints on lower mantle physical properties from seismology and mineral physics. *Earth Planet. Sci. Lett.* **143,** 197–205 (1996).
20. Su, W. & Dziewonski, A. M. Simultaneous inversion for 3-D variations in shear and bulk velocity in the mantle. *Phys. Earth Planet. Inter.* **100,** 135–156 (1997).
21. Kennett, B. L. N., Widiyantoro, S. & van der Hilst, R. D. Joint seismic tomography for bulk-sound and shear wavespeed in the Earth's mantle. *J. Geophys. Res.* **103,** 12469–12493 (1998).
22. Masters, G., Laske, G., Bolton, H. & Dziewonski, A. M. in *Earth's Deep Interior: Mineral Physics and Tomography From the Atomic to the Global Scale.* Geophys. Monogr. Ser. (eds Karato, S.-I. *et al.*) Vol. 117, 63–87 (AGU, Washington, DC, 2000).
23. Ishii, M. & Tromp, J. Normal-mode and free-air gravity constraints on lateral variations in velocity and density of Earth's mantle. *Science* **285,** 1231–1236 (1999).
24. Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25,** 297–356 (1981).
25. Sinelnikov, Y. D., Chen, G. & Liebermann, R. C. Elasticity of $CaTiO_3$-$CaSiO_3$ perovskites. *Phys. Chem. Mineral.* **25,** 515–521 (1998).
26. Zhang, J. & Weidner, D. J. Thermal equation of state of aluminum-enriched silicate perovskite. *Science* **284,** 782–784 (1999).

© 2001 Macmillan Magazines Ltd

27. McDonough, W. F. & Sun, S.-S. The composition of the Earth. *Chem. Geol.* **120,** 223–253 (1995).
28. Ekström, G. & Dziewonski, A. M. The unique anisotropy of the Pacific upper mantle. *Nature* **394,** 168–172 (1998).
29. Karato, S. Importance of anelasticity in the interpretation of seismic tomography. *Geophys. Res. Lett.* **20,** 1623–1626 (1993).
30. Forte, A. M. & Peltier, W. R. The kinematics and dynamics of poloidal-toroidal coupling in mantle flow: The importance of surface plates and lateral viscosity variations. *Adv. Geophys.* **36,** 1–119 (1994).
31. Karato, S.-I. & Li, P. Diffusion creep in perovskite: Implications for the rheology of the lower mantle. *Science* **255,** 1238–1240 (1992).
32. Constable, S. C., Parker, R. L. & Constable, C. G. Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics* **52,** 289–300 (1987).
33. King, S. & Masters, T. G. An inversion for the radial viscosity structure using seismic tomography. *Geophys. Res. Lett.* **19,** 1551–1554 (1992).
34. Mitrovica, J. X. & Forte, A. M. The radial profile of mantle viscosity: Results from the joint inversion of convection and post-glacial rebound observables. *J. Geophys. Res.* **102,** 2751–2769 (1997).
35. Forte, A. M. in *Earth's Deep Interior: Mineral Physics and Tomography From the Atomic to the Global Scale.* Geophys. Monogr. Ser. (eds Karato, S.-I. *et al.*) Vol. 117, 3–36 (AGU, Washington, DC, 2000).
36. Durek, J. & Ekström, G. A radial model of anelasticity consistent with long-period surface-wave attenuation. *Bull. Seismol. Soc. Am.* **86,** 144–158 (1996).
37. Karato, S.-I. A dislocation model of seismic wave attenuation and micro-creep in the Earth. *Pure Appl. Geophys.* **153,** 239–256 (1998).
38. Hofmann, A. W. The gabbro-harzburgite connection in OIB sources. *Eos Trans. AGU* **80** (Fall Meeting suppl.) 1182–1183 (1999).
39. Ottino, J. M. *The Kinematics of Mixing: Stretching, Chaos, and Transport* Ch. 2 & 4 (Cambridge Univ. Press, Cambridge, 1989).
40. Kellogg, J. B. & O'Connell, R. J. The effects of toroidal motion and layered viscosity on mixing in three dimensions. *Eos Trans. AGU* **80** (Fall Meeting suppl.) 949 (1999).
41. Hager, B. H. Subducted slabs and the geoid: Constraints on mantle rheology and flow. *J. Geophys. Res.* **89,** 6003–6015 (1984).
42. Forte, A. M. & Peltier, W. R. Plate tectonics and aspherical Earth structure: The importance of poloidal-toroidal coupling. *J. Geophys. Res.* **92,** 3645–3679 (1987).
43. Ricard, Y. & Vigny, C. Mantle dynamics with induced plate tectonics. *J. Geophys. Res.* **94,** 17543–17559 (1989).
44. Gurnis, M. & Davies, G. F. The effect of depth-dependent viscosity on convective mixing of the mantle and the possible survival of primitive mantle. *Geophys. Res. Lett.* **13,** 541–544 (1986).
45. Manga, M. Mixing of heterogeneities in the mantle—Effects of viscosity differences. *Geophys. Res. Lett.* **23,** 403–406 (1996).
46. van Keken, P. & Ballentine, C. J. Whole-mantle versus layered mantle convection and the role of a high-viscosity lower mantle in terrestrial volatile evolution. *Earth Planet. Sci. Lett.* **156,** 19–32 (1998).
47. Bunge, H.-P. & Richards, M. A. The origin of large-scale structure in mantle convection: Effects of plate motions and viscosity stratification. *Geophys. Res. Lett.* **23,** 2987–2990 (1996).
48. Ritsema, J., van Heijst, H. J. & Woodhouse, J. H. Complex shear velocity structure imaged beneath Africa and Iceland. *Science* **286,** 1925–1928 (1999).
49. Gurnis, M., Mitrovica, J. X., Ritsema, J. & van Heijst, H. J. Constraining mantle density structure using geological evidence of surface uplift rates: The case of the African superplume. *Geochem. Geophys. Geosystems.* [online] **1** (2000).
50. Jackson, I. Elasticity, composition and temperature of the Earth's lower mantle: A reappraisal. *Geophys. J. Int.* **134,** 291–311 (1998).
51. Stacey, F. D. Thermoelasticity of a mineral composite and a reconsideration of lower mantle properties. *Phys. Earth Planet. Inter.* **106,** 219–236 (1998).
52. Jackson, I., Paterson, M. S. & Fitzgerald, J. D. Seismic wave attenuation in Åheim dunite: An experimental study. *Geophys. J. Int.* **108,** 517–534 (1992).
53. Zerr, A., Diegler, A. & Boehler, R. Solidus of Earth's deep mantle. *Science* **281,** 243–246 (1998).
54. Stacey, F. D. Theory of thermal and elastic properties of the lower mantle and core. *Phys. Earth Planet. Inter.* **89,** 219–245 (1995).

Acknowledgements

We thank I. Jackson for a review, for suggestions used in the Methods section, and for the detailed check of the derivatives in Table 1. A.M.F. also thanks A. Davaille for many helpful discussions. We acknowledge support from NSERC, the Canada Foundation for Innovation, the Ontario Innovation Trust, and the Canadian Institute for Advanced Research—Earth Systems Evolution Program.

Correspondence and requests for materials should be addressed to A.M.F. (e-mail: aforte@uwo.ca).

© 2001 Macmillan Magazines Ltd

Distinct roles of nerve and muscle in postsynaptic differentiation of the neuromuscular synapse

Weichun Lin*, Robert W. Burgess†, Bertha Dominguez*, Samuel L. Pfaff*, Joshua R. Sanes† & Kuo-Fen Lee*

* *The Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA*
† *Department of Anatomy and Neurobiology, Washington University Medical School, St Louis, Missouri 63110, USA*

The development of chemical synapses is regulated by interactions between pre- and postsynaptic cells. At the vertebrate skeletal neuromuscular junction, the organization of an acetylcholine receptor (AChR)-rich postsynaptic apparatus has been well studied. Much evidence suggests that the nerve-derived protein agrin activates muscle-specific kinase (MuSK) to cluster AChRs through the synapse-specific cytoplasmic protein rapsyn. But how postsynaptic differentiation is initiated, or why most synapses are restricted to an 'end-plate band' in the middle of the muscle remains unknown. Here we have used genetic methods to address these issues. We report that the initial steps in postsynaptic differentiation and formation of an end-plate band require MuSK and rapsyn, but are not dependent on agrin or the presence of motor axons. In contrast, the subsequent stages of synaptic growth and maintenance require nerve-derived agrin, and a second nerve-derived signal that disperses ectopic postsynaptic apparatus.

Synapses are the cellular determinants of neural circuitry, and therefore fundamental to nervous system function. A cardinal feature of the chemical synapse is the presence of a postsynaptic apparatus containing high concentrations of neurotransmitter receptors closely associated with numerous extracellular, transmembrane and cytoplasmic scaffolding and signalling components. Perhaps best studied in this respect is the aggregation of acetylcholine receptors (AChRs) at postsynaptic sites of the vertebrate skeletal neuromuscular junction (NMJ). One of the central questions in NMJ formation is the contribution of muscle versus nerve in initiating and maintaining postsynaptic differentiation. In the past three decades, many studies have suggested that motor nerves are largely responsible for organizing this aspect of postsynaptic differentiation through two intramuscular mechanisms: clustering of AChRs initially distributed diffusely in the plane of the muscle membrane; and selective transcription of AChR genes by myonuclei associated with synaptic sites[1,2]. Several nerve-derived molecules that activate these pathways have been identified; of these, agrin has been shown to have a critical role *in vivo*[1,2].

Agrin was identified from its ability to induce clustering of AChRs on cultured myotubes[3,4]. Although muscle fibres also express agrin, neural agrin is 1,000 times more active than muscle agrin in inducing AChR clustering *in vitro*[5,6]. In agrin-deficient mice, the number and density of AChR clusters, as well as synapse-specific transcription, are reduced markedly[7,8]. Agrin functions, at least in part, by activating the muscle-specific kinase (MuSK)[9]. AChR clustering and synapse-specific transcription are completely abolished in MuSK-deficient mice[10]. In rapsyn mutants, AChR clustering is abolished; although, synapse-specific transcription is maintained[11]. Together, these genetic studies provide evidence for a pathway in which agrin acts through MuSK and rapsyn to organize postsynaptic differentiation[1,2].

Although innervation is clearly required to maintain postsynaptic specialization, its role in initiating AChR clustering is still unknown. Indeed, some postsynaptic differentiation has been observed in muscles in which innervation has been ablated using neurotoxins[12,13], or gene-targeting as in topoisomerase IIβ mutants[14], suggesting that muscle-intrinsic mechanisms are sufficient to initiate AChR clustering. In both cases, however, the muscle was transiently innervated, leaving open the possibility that a brief contact is sufficient to initiate postsynaptic differentiation. Here we have used genetic methods to determine the role of nerve and muscle in initiating and maintaining the postsynaptic apparatus.

First, we examined whether the initial stages of postsynaptic differentiation are as dependent on the agrin/MuSK/rapsyn pathway as the later stages are. We show that postsynaptic differentiation can be initiated by mechanisms that are not only independent of agrin but also entirely independent of nerve. MuSK and rapsyn, however, are required for these initial events. Second, we examined why synapses are not randomly distributed in muscles but restricted largely to the central regions of muscle fibres, forming an 'end-plate band'. Our results suggest that a muscle-intrinsic pre-pattern may be a primary determinant of the general localization of the end-plate band. Nerve, on the other hand, supplies both positive and negative signals to refine the location of synapse. Together, our studies provide new insight into the interplay of pre- and postsynaptic signals that lead to differentiation of the postsynaptic apparatus.

Aneural clusters of AChRs at initial stages of synaptogenesis

Previous studies of developing mammalian NMJs emphasized the precise apposition of nerve terminals to AChR-rich postsynaptic sites[1,2]. Motor axons innervate mouse muscles by embryonic day (E)-12.5 and form an intramuscular nerve trunk in the centre of the muscle. Few or no AChR clusters are present at that time, as detected with rhodamine or Texas-red α-bungarotoxin (rBTX), which binds specifically to AChRs[15–18]. By E14.5, however, numerous AChR-rich patches are detectable, most of which are present in an end-plate band near the centre of the muscle (Fig. 1a).

To examine the relationship of these patches to newly formed synapses, we double stained whole-mount diaphragm muscles with rBTX and antibodies to neuronal antigens—either synaptophysin, a synaptic vesicle protein, or neurofilament, a component of the axonal cytoskeleton. As expected, the AChR clusters were near the intramuscular nerve. However, examination at high power revealed that a significant fraction of the AChR clusters were not closely apposed by nerve terminals (Fig. 1a); that is, they were 'aneural clusters'. In contrast, by E16.5–18.5, all AChR clusters were apposed by nerve terminals (Fig. 1b).

We considered that the aneural appearance of AChR clusters might be an artefact if the tips of young axons contained low levels

© 2001 Macmillan Magazines Ltd

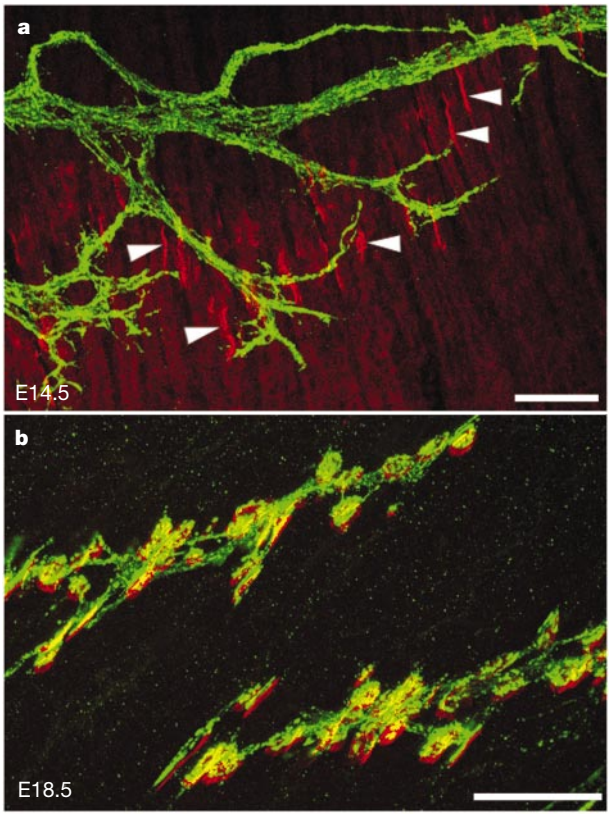


Figure 1 Aneural AChR clusters at initial stage of synaptogenesis. Diaphragm muscles from E14.5 (**a**) and E18.5 (**b**) control embryos were immunostained with anti-synaptophysin (green) plus Texas-red-conjugated α-bungarotoxin (red). The images show that some AChR clusters (red) were not apposed by nerve terminals (green) at E14.5 (aneural AChR clusters, arrowheads in **a**). In contrast, all AChR clusters were apposed by nerve terminals at E18.5 (yellow in **b**). Scale bars, 50 μm.

of synaptic vesicles or neurofilaments. We therefore studied transgenic mice expressing yellow fluorescent protein (YFP) in motoneurons under the control of a Thy1 promoter. We found that motor axons innervating the diaphragm of the Thy1-YFP-16 line are completely labelled by E13.5 (ref. 17), yet many AChR clusters in these mice were not apposed by YFP-labelled axons at E14.5 (data not shown). With YFP labelling, as with synaptophysin or neurofilament labelling, all AChR clusters were innervated by E16.5 (data not shown). From these results, we conclude that some myotubes bear aneural AChR clusters during the initial stages of synaptogenesis, consistent with previous reports[15,19].

Agrin-independent initiation of postsynaptic differentiation

How might aneural AChR clusters arise? Because all clusters were concentrated in the end-plate band, we wondered whether nerve-derived agrin was poorly bound to matrix at this stage, and therefore able to diffuse an appreciable distance from axons. Alternatively, axons might be growing and retracting rapidly, and therefore able to deposit agrin and then withdraw from the resulting AChR clusters. To test these ideas, we examined mutant mice in which we had deleted the 'Z' exons of the agrin gene, which are selectively included in neuronal agrin and are required for maximal agrin bioactivity[8].

We showed previously that neuromuscular synaptogenesis is markedly disrupted in z-agrin-deficient mutants (called AGZ) at E18.5; the number and size of AChR clusters in AGZ mutants at this stage are much lower than those in controls[8]. When we examined E14.5 embryos, however, we found end-plate bands containing a similar number of AChR clusters in AGZ mutants and controls (Fig. 2g, h), although pre-synaptic axons innervated a broader region of muscle in AGZ mutants than in controls (Fig. 2a, b). In addition, the average size and area of individual AChR clusters were similar between genotypes (Table 1). Whereas AChR clusters in controls became larger and more numerous over the next few days, however, those in AGZ mutants declined in both number and size. Thus, while fully consistent with previous observations[8], these results also show that postsynaptic specialization is initiated but not maintained in the absence of z-agrin.

To assess the distribution of other synaptic proteins in AGZ mutant muscles, we examined the location of acetylcholinesterase (AChE) and rapsyn. AChE is a component of the specialized basal lamina that overlies AChR clusters, and, as noted above, rapsyn is a synapse-specific cytoplasmic protein. Aggregates of AChE and rapsyn were present in control and AGZ mutant muscles at E14.5. Both AChE and rapsyn clusters became larger and more numerous in controls by E18.5, but became smaller and far less numerous in mutants (data not shown). The agrin-independence of postsynaptic differentiation is not, therefore, a unique property of AChRs.

Agrin is synthesized by myotubes and Schwann (glia) cells as well as by motoneurons[20–23]. As AGZ mutants retain near-normal levels of z-deleted agrin, we considered whether agrin derived from muscle or nerve-associated Schwann cells might initiate AChR clustering. Because the originally described agrin mutant is a hypomorphic allele that expresses low levels of muscle agrin[7], we were unable to use these mice to test this idea. Therefore we generated a new agrin mutant allele (AGD) in which most of the coding exons were deleted, so that mice completely lacked both neural and non-neural agrin isoforms (see Methods). AChR, AChE and rapsyn clustering in the AGD mutants was similar to that in the AGZ mutant mice at both E14.5 and E18.5 (Fig. 2i, l, Table 1; and data not shown). Similar to controls (Fig. 2m, arrowheads), some AChR clusters were not apposed by nerve terminals in E14.5 AGZ and AGD mutants (Fig. 2n, o, arrowheads). Thus, neither neural nor non-neural agrin is required to initiate postsynaptic differentiation.

Nerve-independent initiation of postsynaptic development

If agrin is dispensable for initiating postsynaptic differentiation in the end-plate band, the nerve might provide a second organizing signal or the muscle might have an intrinsic capacity to initiate postsynaptic specializations. Indeed, myotubes cultured in the absence of neurons can form clusters of AChRs[24]. To determine whether initiation of postsynaptic development is independent of nerve, we examined mutants lacking the motoneuron-specific transcription factor HB9, in which the phrenic nerve fails to form, so the diaphragm muscle develops without innervation[25,26].

We confirmed that no motor axons were present in the dia-

Table 1 Analysis of AChR clusters in the control, AGZ, AGD and HB9 mutant embryos

	Length (μm)		Area (μm^2)	
	E14.5	E18.5	E14.5	E18.5
Control	11.67 ± 0.80 (48)	15.20 ± 0.39 (99)	11.15 ± 1.03 (48)	59.85 ± 3.15 (99)
AGZ	10.17 ± 0.43* (68)	4.13 ± 0.34† (38)	10.04 ± 0.48*(68)	9.27 ± 1.23† (38)
AGD	12.00 ± 0.67*(37)	4.77 ± 0.39†(44)	13.69 ± 1.26* (37)	10.81 ± 1.17†(44)
HB9	7.07 ± 0.40† (53)	10.19 ± 0.28† (92)	5.66 ± 0.61†(53)	65.82 ± 3.02‡ (92)

Length and area of individual AChR clusters were measured as described in Methods. Values are mean ± s.e.m.; number in parentheses indicates number of AChR clusters measured. Compared with control: * $P > 0.05$, not significant; † $P < 0.0001$, significant; ‡ $P > 0.5$, not significant.

© 2001 Macmillan Magazines Ltd

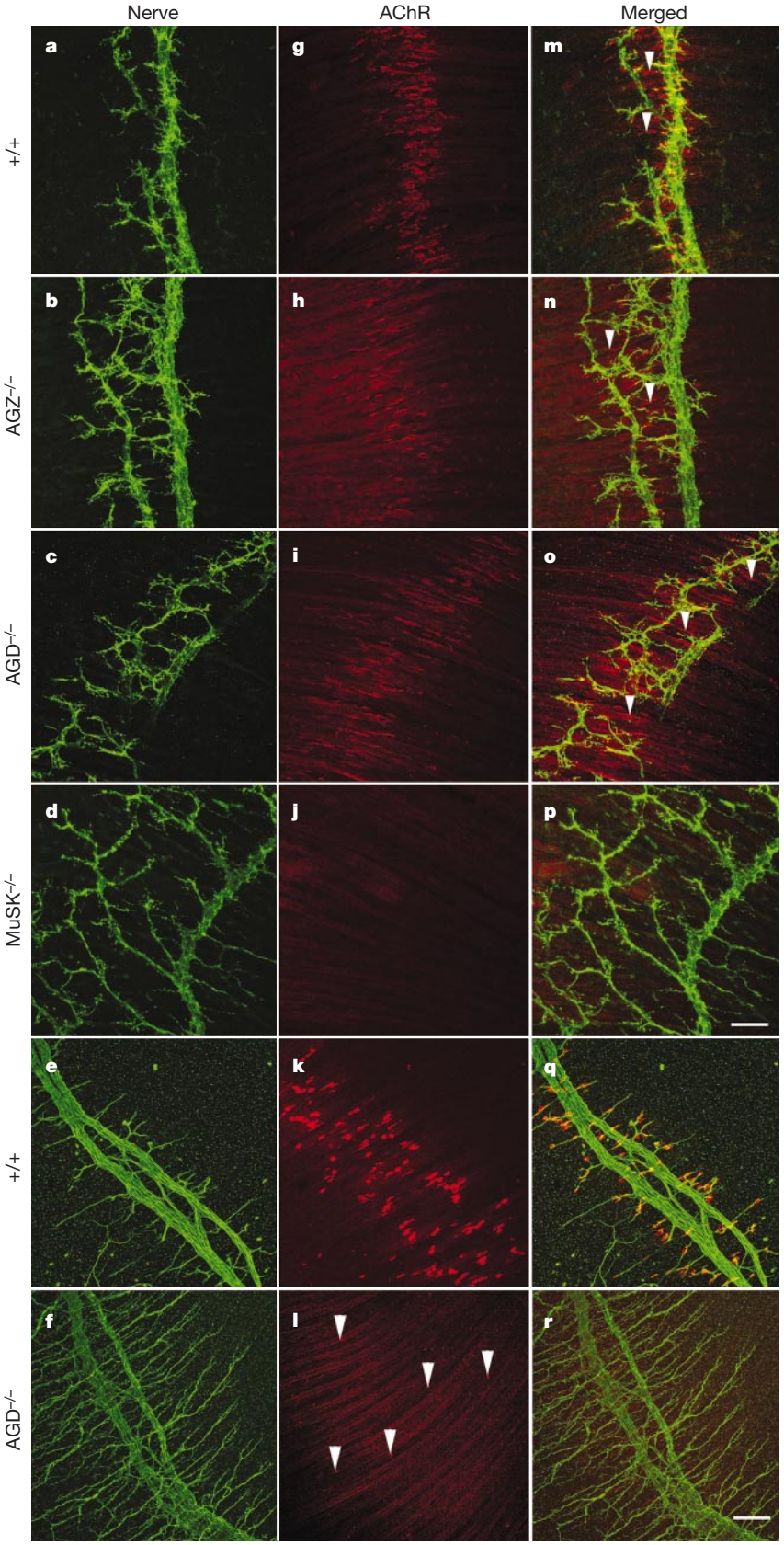


Figure 2 Distinct roles of agrin and MuSK in initiating AChR clustering. Diaphragm muscles from E14.5 (**a**–**d**, **g**–**j**, **m**–**p**) or E18.5 (**e**, **f**, **k**, **l**, **q**, **r**) controls (**a**, **e**), AGZ (**b**), AGD (**c**, **f**), or MuSK (**d**) mutants were immunostained with anti-synaptophysin (**a**–**d**) or anti-neurofilament (**e**, **f**) plus Texas-red-conjugated α-bungarotoxin (**g**–**l**). Numerous AChR clusters were detected in AGZ (**h**) and AGD (**i**) mutants at E14.5. Merged images (**m**–**r**) show that some AChR clusters were not apposed by nerve terminals (arrowheads) at this stage. In contrast, the E14.5 MuSK mutants displayed no AChR clustering (**j**). In addition, the phrenic nerve innervated a broader area of muscle in MuSK mutants (**d**) than in AGZ (**b**) and AGD (**c**) mutants. At E18.5, all AChR clusters were apposed by nerve terminals in the controls (**q**). Only few smaller and dimmer AChR clusters remained in AGD mutants (**l**, arrowheads). Pre-synaptic axons grow extensively in E18.5 AGD mutants (**f**). Scale bars, 50 μm (**a**–**d**, **g**–**j**, **m**–**p**); 100 μm (**e**, **f**, **k**, **l**, **q**, **r**).

© 2001 Macmillan Magazines Ltd

phragm of HB9 mutants from E12.5 to E18.5 (Fig. 3b, d; and data not shown). Nonetheless, AChR clusters were not only present in E14.5 HB9 mutant diaphragms, but also formed an end-plate band in the centre of the muscle. The distribution of clusters was similar in mutants and controls (Fig. 3f, g), although clusters were significantly smaller in the former (Table 1). Thus, early stages of postsynaptic differentiation can occur in the presumptive end-plate band even in the absence of nerves, presumably by mechanisms intrinsic to muscle.

On the basis of these results, we examined whether other aspects of the postsynaptic apparatus develop in the absence of motor nerves. To determine whether nuclei in the end-plate band become transcriptionally specialized, we used *in situ* hybridization to assess the distribution of AChR α-subunit messenger RNA in the diaphragm. We found that AChRα transcripts were concentrated in the central region of the diaphragm muscle of E15.5 controls (Fig. 4a) and HB9 mutants (Fig. 4e). Likewise, aggregates of AChE were present in end-plate bands of both control (Fig. 4c) and HB9 mutant (Fig. 4g) muscles. For both markers, the pattern was more diffuse in HB9 mutants than in controls. Clearly, however, the muscle can organize and coordinate multiple aspects of differentiation in the absence of the nerve.

Positive and negative effects of axons

Although AChR and AChE clusters in aneural and agrin-deficient muscles were similar in many respects at early stages, they differed markedly at later stages. In innervated but agrin-deficient muscles, AChR clusters became smaller, dimmer and much less numerous as development proceeded (Fig. 2l). In contrast, clusters remained numerous and increased in size between E14.5 and E18.5 in the aneural muscles in HB9 mutants (Fig. 3i and Table 1). Moreover, central concentrations of AChR mRNA and aggregates of AChE remained prominent in HB9 mutant muscles at E18.5 (Fig. 4f, h), whereas, consistent with previous results[7,8], neither feature was present in E18.5 AGD or AGZ mutant muscle (data not shown).

Some differences between clusters in HB9 mutant and wild-type muscles were apparent at E18.5: the end-plate band was more than twice the width in HB9 mutant mice compared with in controls (345.29 ± 19.84 μm, $n = 17$, versus 124.44 ± 10.68 μm, $n = 13$; $P < 0.005$), and individual clusters were shorter and rounder in mutants than in controls at all ages (Table 1; Fig. 5a, b). Nonetheless, postsynaptic differentiation proceeded to a far greater extent in the absence of the entire nerve than in the absence of only agrin.

A possible explanation for this apparent paradox is that muscle agrin might maintain postsynaptic differentiation in the absence of nerves, even though it is unable to do so in the absence of neural agrin. To test this admittedly unlikely possibility, we generated HB9/AGD double mutants. We found that similar numbers of AChR clusters were present in HB9 single mutant (Fig. 3i) and HB9/AGD double-mutant muscles (Fig. 3j), whereas few clusters persisted at E18.5 in AGD single mutants (Fig. 2l). Together, these results indicate that the nerve supplies muscle not only with agrin to maintain AChR clusters but also with an agrin-independent signal that disperses postsynaptic specializations that have not been stabilized by nerve-derived agrin.

MuSK and rapsyn are required for initiation of synaptogenesis

As MuSK and rapsyn are critical components of the AChR clustering machinery, we examined whether they are also required for the early stages of postsynaptic differentiation. No AChR clusters were detectable in MuSK or rapsyn mutants at E14.5 (Fig. 2j; and data not shown), as has been shown to be the case at later stages[10,11]. These results suggest that initiation of postsynaptic differentiation is dependent on both MuSK and rapsyn. Notably, excess axon growth at E14.5 was much more prominent in MuSK mutants (Fig. 2d) than in AGZ, AGD and rapsyn mutants (Fig. 2b, c; and data not shown). These results show distinct roles for agrin and MuSK

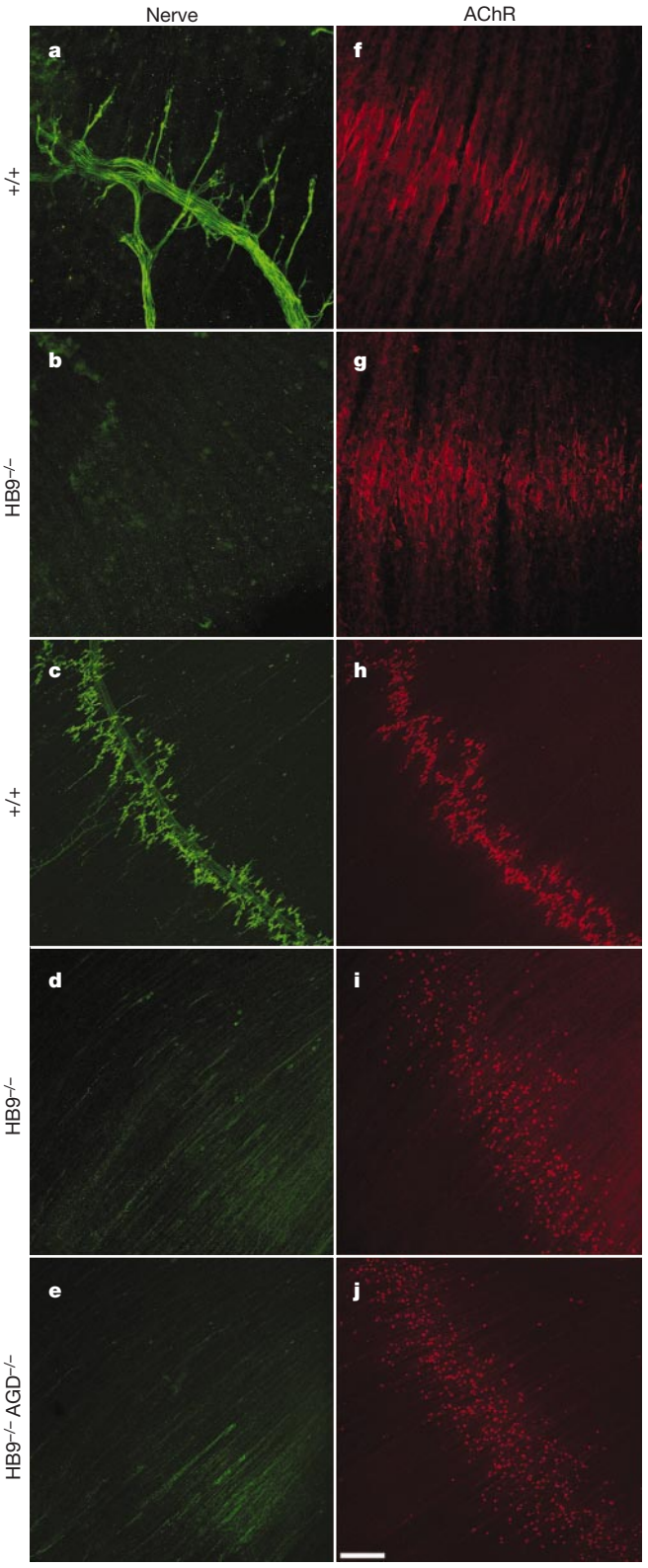


Figure 3 Clustering of AChRs in the absence of the motor nerve. Diaphragm muscles from E14.5 (**a**, **b**) or E18.5 (**c**–**e**) control (**a**, **c**), HB9$^{-/-}$ (**b**, **d**) or HB9/AGD double mutants (**e**, HB9$^{-/-}$AGD$^{-/-}$) were immunostained with antibodies to neurofilament (**a**, **b**) or synaptophysin (**c**–**e**) plus Texas-red-conjugated α-bungarotoxin (**f**–**j**). Clustering of AChRs was observed in the presumptive end-plate band of HB9 mutants at E14.5 (**g**) in the absence of innervation (**b**). AChR clusters were maintained as development proceeded in HB9 mutants (**i**). Similarly, AChR clusters were maintained in HB9/AGD double mutants (**j**). Scale bars, 50 μm (**a**, **b**, **f**, **g**); 200 μm (**c**–**e**, **h**–**j**).

© 2001 Macmillan Magazines Ltd

during the initial stages of NMJ development.

To determine whether nerve-independent aspects of postsynaptic differentiation are dependent on MuSK, we generated HB9/MuSK double mutants to produce aneural, MuSK-deficient muscles. We found that AChR clustering was completely absent in the HB9/MuSK double mutants (Fig. 5e), as it was in the MuSK single mutants (Fig. 5c). Thus, MuSK is required for nerve-independent AChR clustering. Again, this requirement might reflect the existence of an alternate, muscle-derived ligand for MuSK. For example, the extracellular domain of MuSK bears similarities to the Wnt receptor Frizzled[27,28], raising the possibility that Wnt-like ligands, which have been implicated in central synaptogenesis[29], act through MuSK to initiate postsynaptic differentiation. Alternatively, cluster formation might result from spontaneous, ligand-independent activation of MuSK.

Our previous structure–function analysis of MuSK provided evidence for ligand-independent activation[30], and our observations in the heterozygous mutants provide indirect support for this possibility *in vivo*. In $MuSK^{+/-}$ embryos (Fig. 5d), AChR clustering was similar to that in controls (Fig. 5a). In contrast, the number of AChR clusters in $HB9^{-/-}MuSK^{+/-}$ embryos was less than ~5% of those observed in the controls (Fig. 5, compare f with a or d). Thus, levels of MuSK are not limiting when nerve is present, but are limiting when nerve is absent. This result is expected if autoactivation, which should be a steep function of MuSK concentration, initiates postsynaptic differentiation in the absence of agrin.

Discussion

We have analysed early stages of postsynaptic differentiation in muscles of mutant mice lacking agrin, MuSK, rapsyn and/or motor nerves (Table 2). Our results are consistent with previous reports (refs 7, 8, 10–14, 19, 25, 26, 33, 34) but help reconcile seemingly discrepant views on the nerve dependence of postsynaptic differentiation. For example, although similar postsynaptic defects are observed in MuSK and agrin mutants at birth[7,10], our results suggest that the defect in MuSK mutants is due to an absence of initiation of postsynaptic differentiation, whereas the impairment in agrin mutants is caused by a loss of agrin-dependent maintenance of

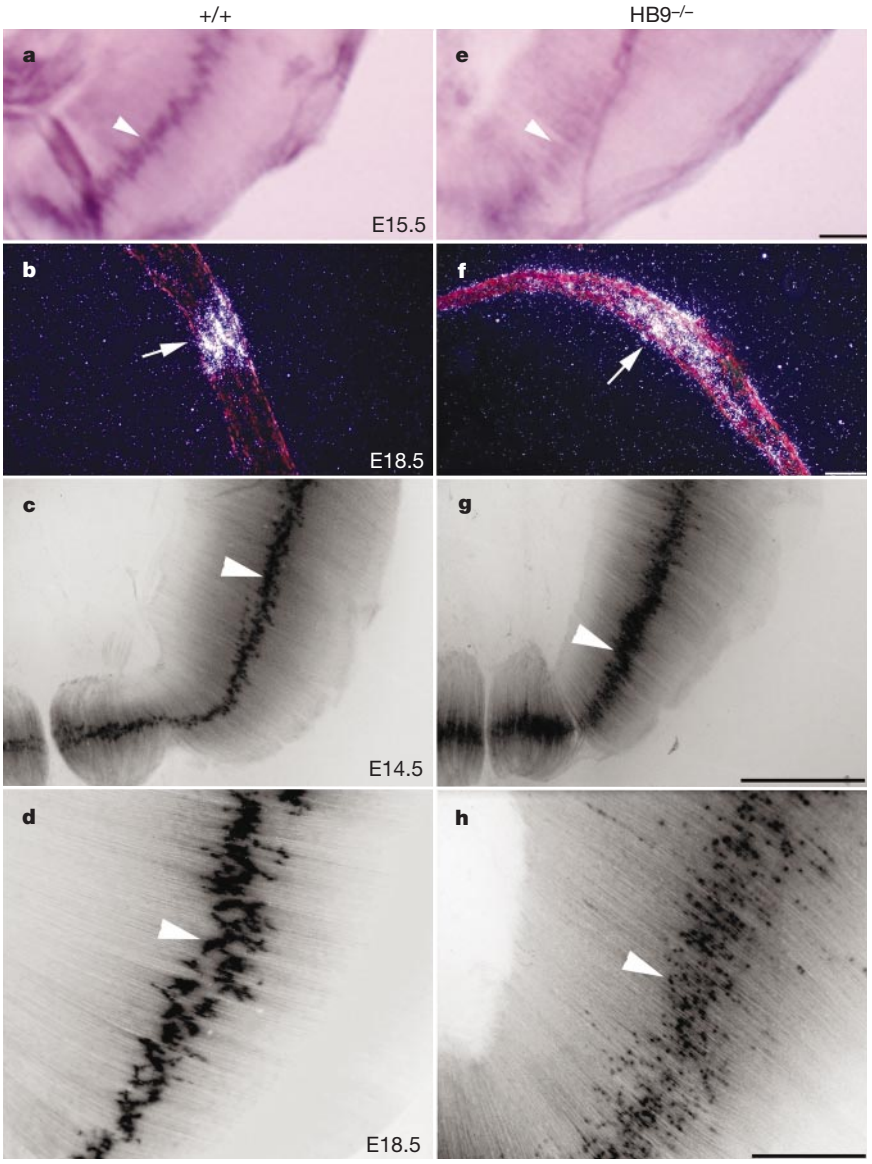


Figure 4 Localization of AChRα mRNA and AChE in the presumptive end-plate band of aneural muscle. *In situ* hybridization of E15.5 whole-mount diaphragm muscles (**a**, **e**) or longitudinal sections (**b**, **f**) of E18.5 diaphragm muscles from controls (**a**, **b**) and mutants (**e**, **f**) showed that AChRα transcripts were concentrated in the central region of muscle (arrowheads in **a** and **e**; arrows in **b** and **f**). Similarly, AChE was also clustered in control (**c**, **d**, arrowheads) and aneural (**g**, **h**, arrowheads) diaphragm muscles. Scale bars, 200 μm (**a**, **e**); 100 μm (**b**, **f**); 1 mm (**c**–**g**); 500 μm (**d**, **h**).

© 2001 Macmillan Magazines Ltd

the postsynaptic apparatus. On the basis of these and previous studies, we propose the existence of three early, overlapping steps in the formation of the postsynaptic apparatus at the NMJ (Fig. 6). First, a muscle-intrinsic, nerve/agrin-independent and MuSK-dependent mechanism initiates formation of postsynaptic specialization in an end-plate band. Second, nerve-derived agrin acts through MuSK to promote apposition of nerve terminals to these nerve-independent AChR clusters and/or to induce new postsynaptic sites. Agrin is also required for the growth and maintenance of most, if not all, synaptic sites. Third, motor axons, or Schwann cells that accompany them, provide an agrin-independent signal that destabilizes or disperses postsynaptic apparatus that have not been stabilized by agrin.

This scheme provides a framework for asking new questions. First, what determines the location of the end-plate band? It was previously supposed that this region is determined by the position at which axons enter the muscle, but this is clearly not the case. An alternative possibility is suggested by the haplo-insufficiency of MuSK as initiator of nerve-independent AChR clustering. As myotubes generally grow from the centre toward the ends[31], central nuclei may be the first to differentiate. If these central nuclei are the first to express the MuSK gene, the resulting local synthesis of MuSK might make the central band the preferred site of postsynaptic differentiation. Indeed, it is possible that other aspects of this central differentiation lead to selective localization of determinants that help promote formation of the intramuscular nerve in the muscle's centre. In this scheme, muscle rather than nerve would be the primary determinant of the general localization of synaptic sites, although nerve might still be responsible for the precise location of individual postsynaptic apparatus.

Second, what role does nerve-independent postsynaptic differentiation have in normal synaptogenesis? Studies on adult muscle show that regenerating axons selectively innervate pre-existing sites[32], whereas studies of cultured muscle show that axons can organize new AChR clusters at sites of contact, and disperse pre-existing clusters[33,34]. Either of these models might apply to developing muscle *in vivo*. On the one hand, axons might be attracted to pre-existing clusters or encounter them by chance, and then use agrin to stabilize or enlarge them. Alternatively, nerve-independent clusters might resemble "spontaneous hot spots" in culture, and

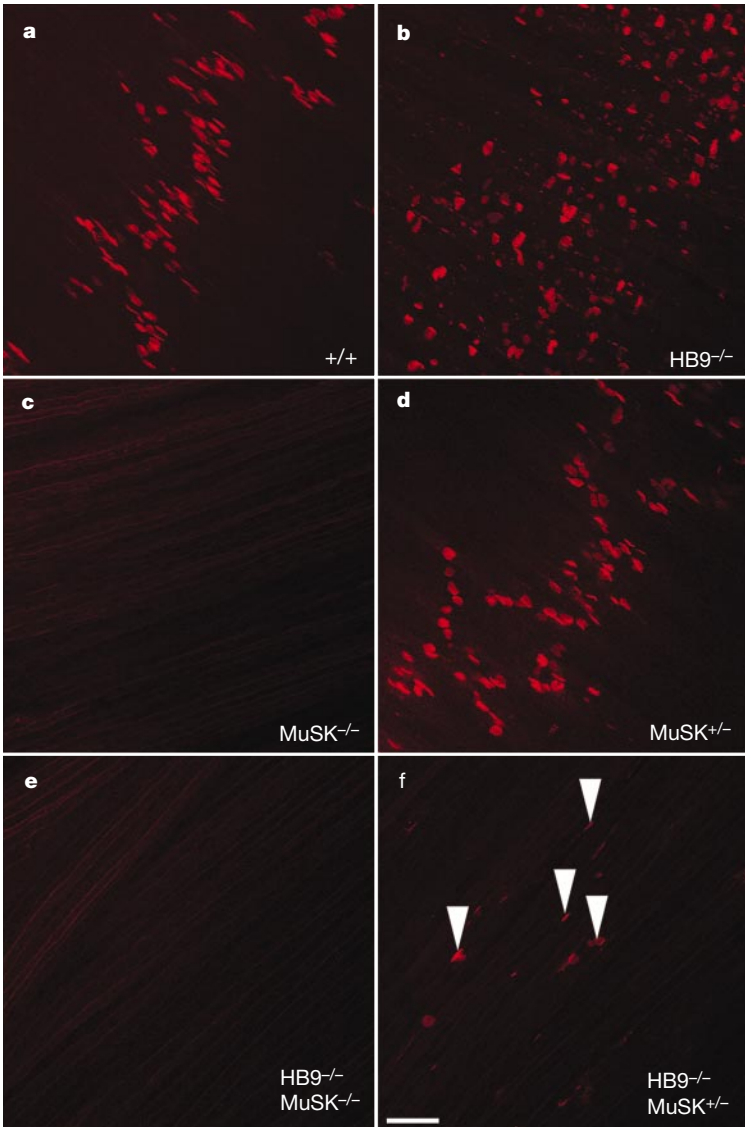


Figure 5 MuSK is required for AChR clustering in aneural muscle. Diaphragm muscles from E18.5 mutants of the indicated genotypes were labelled with Texas-red-conjugated α-bungarotoxin. In contrast to abundant AChR clusters observed in HB9 single mutants (**b**), AChR clustering was absent in the HB9/MuSK double mutants (**e**, HB9$^{-/-}$MuSK$^{-/-}$), as in the MuSK single mutants (**c**, MuSK$^{-/-}$). Unexpectedly, AChR clustering was reduced markedly in embryos heterozygous for MuSK and homozygous for HB9 alleles (**f**, HB9$^{-/-}$ MuSK$^{+/-}$). In contrast, MuSK heterozygous (MuSK$^{+/-}$) mutants displayed abundant AChR clusters (**d**), similar to those observed in controls (**a**). Scale bar, 50 μm.

© 2001 Macmillan Magazines Ltd

Table 2 Summary of genetic analysis of AChR clustering in the control and mutant embryos at E14.5 and E18.5 stages

Genotype	AChR clusters (aneural; A)*	
	E14.5	E18.5
Controls	Many (some A)	Many (no A)
$AGZ^{-/-}$	Many (some A)	Few (most A)
$AGD^{-/-}$	Many (some A)	Few (most A)
$HB9^{-/-}$	Many (all A)	Many (all A)
$MuSK^{+/-}$	Many (some A)	Many (no A)
$MuSK^{-/-}$	None (–)	None (–)
$HB9^{-/-}AGZ^{-/-}$	Many (all A)	Many (all A)
$HB9^{-/-}AGD^{-/-}$	Many (all A)	Many (all A)
$HB9^{-/-}MuSK^{+/-}$	ND	Few (all A)
$HB9^{-/-}MuSK^{-/-}$	ND	None (–)

* A, aneural AChR clusters that are not in close apposition with nerve terminals; ND, not determined.

axons might use agrin to organize new postsynaptic sites. Even in this case, however, activation of preexisting AChR clusters by ACh released from incoming motor axons[35,36] may render the central band of muscle a receptive environment and other regions of muscle non-receptive, so that motor end plates are confined to a central region of muscle fibres. Indeed, even though synaptic sites bear stop signals for axonal growth[2], the excessive axon growth in agrin and MuSK mutant mice may, in part, be caused by the loss or absence of AChR clusters *per se*. Consistent with this possibility, excess axon growth is more prominent in MuSK than in agrin mutants at E14.5 (Fig. 2b–d), whereas excess axon growth becomes more prominent in the agrin mutants at later stages, coincident with dispersion of AChR clusters (for example, Fig. 2l). A third alternative is suggested by the observation that no AChR clusters are detected in wild-type muscle before axons arrive. Muscle-intrinsic cues might serve as a redundant mechanism to enhance synaptic receptivity, ensuring that all myotubes do become innervated.

Finally, what is the agrin-independent neural signal that disperses ectopic postsynaptic sites? One obvious possibility is that nerve-evoked electrical activity, already known to repress AChR gene transcription in non-synaptic nuclei[37,38], leads to secondary loss of aneural AChR clusters. There is also evidence, however, that nerves can exert shorter range, activity-independent repressive effects of AChR synthesis or clustering[39–41]. A balance of nerve-derived positive and negative signals with different spatial ranges could enable NMJs to develop at the central band of muscle fibres, thereby facilitating the temporal and spatial coordination of contractile activation required for movement. □

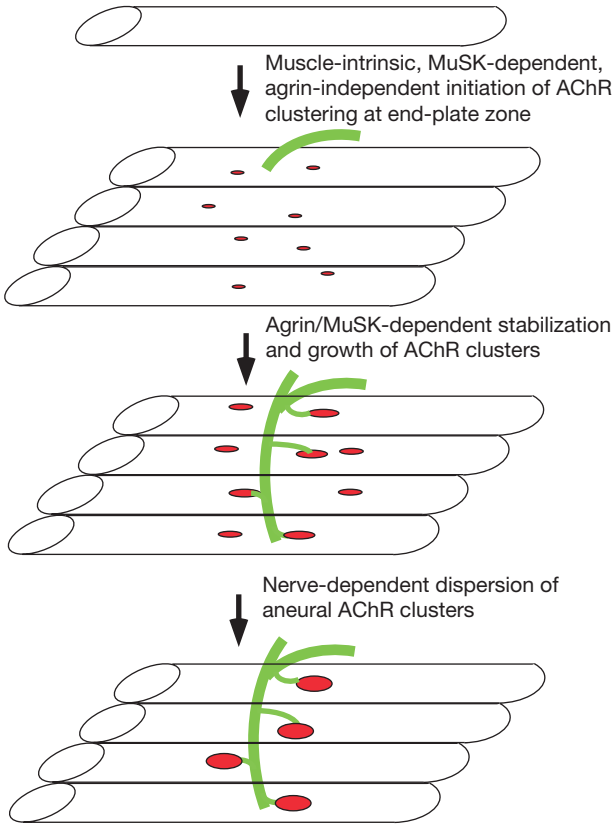


Figure 6 Multiple steps in postsynaptic differentiation in the neuromuscular junction. Initiation of AChR clustering is determined by a muscle-intrinsic mechanism that is dependent on MuSK and independent at the agrin. This step is followed by two overlapping steps in which AChR clusters are formed or stabilized by neural agrin, whereas aneural AChR clusters are dispersed by the nerve through an agrin-independent mechanism, leading to a close apposition of nerve terminals and AChR clusters. White, muscle fibre; red, AChR clusters; Green, nerve.

Methods

Animals

The use of animals is in compliance with the guidelines of Institute Animal Care and Use Committee of the Salk Institute. To generate agrin-null mice (AGD for deleted agrin), we cloned the 4.5-kilobase (kb) *Not*I/*Bam*HI genomic fragment[42] into the pLoxPneo vector upstream of the PGK-Neo cassette. A second genomic fragment extending from intron 33 to exon 36 was generated by polymerase chain reaction with *Sal*I/*Xho*I ends, and cloned into the *Xho*I site of the same vector. A thymidine kinase gene was then ligated into the *Sac*I/*Not*I sites of pLoxPneo. This allele deletes ~9 kb of DNA, including most of the agrin coding region, from the *Bam*HI site in exon 6 to downstream of exon 33 (the second Z exon). This construct was electroporated into embryonic stem cells and a homologous recombinant clone was injected into blastocysts. We bred the resulting chimaeric mice to generate heterozygous and eventually homozygous animals. Homologous recombination was verified by Southern analysis, and complete absence of agrin was verified by immunohistochemistry and immunoblotting. We carried out genotyping using an antisense primer in intron 33 and a primer in the 3′ end of PGK-Neo to identify the mutant allele, and a primer in exon 31 and an antisense primer in exon 32 to identify wild-type alleles.

Thy-1-YFP[17], HB9[26], AGZ[8] and MuSK[10] mutants have been described. Standard breeding methods were used to generate double mutants.

Immunohistochemistry

Diaphragm muscles were fixed in 2% paraformaldehyde (PFA) in 0.1 M phosphate buffer (pH 7.3) at 4 °C overnight, rinsed briefly with phosphate-buffered saline (PBS, pH 7.3), incubated in 0.1 M glycine in PBS for 1 h, and rinsed briefly with PBS followed by 0.5% Triton X-100 in PBS. The muscles were blocked in dilution buffer (500 mM NaCl, 0.01 M phosphate buffer, 3% bovine serum albumin, 5% goat serum and 0.01% thimerosal) overnight at 4 °C, and then incubated with primary rabbit antibodies against neurofilament 150 (Chemicon; 1:1,000) or synaptophysin (a gift from P. De Camilli; 1:1,000) in dilution buffer overnight at 4 °C. After three 1-h washes in 0.5% Triton X-100 in PBS, the muscles were incubated with fluorescein-conjugated goat anti-rabbit IgG (1:400, Cappel) and Texas-red-conjugated α-BTX (10^{-8} M; Molecular Probes) overnight at 4 °C. We then washed the muscles three times for 1 h with 0.5% Triton X-100 in PBS, once with PBS, and flat-mounted them in solution containing 0.5% *N*-propyl gallate and 90% glycerol in 10 mM Tris buffer (pH 8.5). Images were collected with an Olympus confocal microscope.

Quantitative analysis of AChR clusters

Z-serial images were collected from whole-mount diaphragm samples with a ×60 oil objective using an Olympus confocal laser scanning microscope. A single projected image was created by overlaying each set of *z* series. We traced individual AChR clusters from projected images around their perimeters, and calculated the area and length (the longest axis) of individual clusters using Fluoview (version 2.0) software supplied by Olympus.

AChE staining

Briefly, tissues were dissected from embryos fixed with 2% PFA, rinsed in TBS several times, and incubated with solution containing ethopropazine (0.2 mM), acetylthiocholine iodine (4 mM), glycine (10 mM), cupric sulphate (2 mM) and sodium acetate (65 mM, pH 5.5) for 2–4 h at 37 °C. Staining for AChE was developed by incubating the tissues for 2–5 min in sodium sulphide (1.25%, pH 6.5)[43]. The tissues were then washed extensively with water, cleared with 50% glycerol in PBS, and mounted on a glass slide before photographing.

In situ hybridization

For whole-mount *in situ* hybridization, we fixed diaphragm muscles in 2% PFA in 0.1 M phosphate buffer at 4 °C overnight. A digoxygenin-labelled AChRα riboprobe was transcribed *in vitro*. We carried out hybridization at 70 °C overnight in hybridization buffer (50% formamide, 1.3× SSC, 5 mM EDTA, 50 μg ml^{-1} yeast transfer RNA, 0.2%

© 2001 Macmillan Magazines Ltd

Tween-20, 0.5% CHAPS and 100 μg ml^{-1} heparin). After hybridization, the samples were washed three times for 1 h with TBS containing 1% Tween-20, blocked with 5% goat serum in dilution buffer, and incubated with alkaline phosphatase conjugated anti-digoxygenin (1:1,000; Boehringer Mannheim) overnight at 4 °C. Detection was carried out in 100 mM Tris (pH 9.5) containing nitroblue tetrazolium chloride and 5-bromo-4-chloro-3-indolyl-phosphate (NBT/BCIP). For *in situ* hybridization with a ^{33}P-labelled riboprobe, diaphragm sections (20 μm) were hybridized with a ^{33}P-labelled AChRα riboprobe. Slides were dipped in NTB2 emulsion, exposed for 5 days, photographically processed and counterstained with eosin/haematoxylin.

Received 13 December 2000; accepted 26 February 2001.

1. Burden, S. J. The formation of neuromuscular synapses. *Genes Dev.* **12,** 133–148 (1998).
2. Sanes, J. R. & Lichtman, J. W. Development of the vertebrate neuromuscular junction. *Annu. Rev. Neurosci.* **22,** 389–442 (1999).
3. Nitkin, R. M. *et al.* Identification of agrin, a synaptic organizing protein from Torpedo electric organ. *J. Cell Biol.* **105,** 2471–2478 (1987).
4. McMahan, U. J. The agrin hypothesis. *Cold Spring Harb. Symp. Quant. Biol.* **55,** 407–418 (1990).
5. Ferns, M. *et al.* RNA splicing regulates agrin-mediated acetylcholine receptor clustering activity on cultured myotubes. *Neuron* **8,** 1079–1086 (1992).
6. Ruegg, M. A. *et al.* The agrin gene codes for a family of basal lamina proteins that differ in function and distribution. *Neuron* **8,** 691–699 (1992).
7. Gautam, M. *et al.* Defective neuromuscular synaptogenesis in agrin-deficient mutant mice. *Cell* **85,** 525–535 (1996).
8. Burgess, R., Nguyen, Q. T., Son, Y.-J., Lichtman, J. W. & Sanes, J. R. Alternatively spliced isoforms of nerve- and muscle-derived agrin: their roles at the neuromuscular junction. *Neuron* **23,** 33–44 (1999).
9. Glass, D. J. *et al.* Kinase domain of the muscle-specific receptor tyrosine kinase (MuSK) is sufficient for phosphorylation but not clustering of acetylcholine receptors: required role for the MuSK ecto-domain? *Proc. Natl Acad. Sci. USA* **94,** 8848–8853 (1997).
10. DeChiara, T. M. *et al.* The receptor tyrosine kinase MuSK is required for neuromuscular junction formation *in vivo*. *Cell* **85,** 510–512 (1996).
11. Gautam, M. *et al.* Failure of postsynaptic specialization to develop at neuromuscular junctions of rapsyn-deficient mice. *Nature* **377,** 232–236 (1995).
12. Braithwaite, A. W. & Harris, A. J. Neural influence on acetylcholine receptor clusters in embryonic development of skeletal muscles. *Nature* **279,** 549–551 (1979).
13. Harris, A. J. Embryonic growth and innervation of rat skeletal muscles. III. Neural regulation of junctional and extra-junctional acetylcholine receptor clusters. *R. Soc. Phil. Trans. B* **293,** 287–314 (1981).
14. Yang, X., Li, W., Prescott, E. D., Burden, S. J. & Wang, J. C. DNA topoisomerase IIβ and neural development. *Science* **287,** 131–134 (2000).
15. Lupa, M. T. & Hall, Z. W. Progressive restriction of synaptic vesicle protein to the nerve terminal during development of the neuromuscular junction. *J. Neurosci.* **9,** 3937–3945 (1989).
16. Morris, J. K. *et al.* Genetic rescue of cardiac defect in erbB2 null mutant mice reveals essential roles of erbB2 in development of the peripheral nervous system. *Neuron* **23,** 273–283 (1999).
17. Feng, G. *et al.* Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. *Neuron* **28,** 41–51 (2000).
18. Lin, W. *et al.* Aberrant development of motor axons and synapses in erbB2-deficient mice. *Proc. Natl Acad. Sci. USA* **97,** 1299–1304 (2000).
19. Dahm, L. M. & Landmesser, L. T. The regulation of synaptogenesis during normal development and following activity blockade. *J. Neurosci.* **11,** 238–255 (1991).
20. Godfrey, E. W. *et al.* Basal lamina components are concentrated in premuscle masses and at early acetylcholine receptor clusters in chick embryo hindlimb muscles. *Dev. Biol.* **130,** 471–486 (1988).
21. Fallon, J. R. & Gelfman, C. E. Agrin-related molecules are concentrated at acetylcholine receptor clusters in normal and aneural developing muscle. *J. Cell Biol.* **108,** 1527–1535 (1989).
22. Hoch, W., Ferns, M., Campanelli, J. T., Hall, Z. W. & Scheller, R. H. Developmental regulation of highly active alternatively spliced forms of agrin. *Neuron* **11,** 479–490 (1993).
23. Halfter, W. *et al.* Distribution and substrate properties of agrin, a heparan sulfate proteoglycan of developing axonal pathways. *J. Comp. Neurol.* **383,** 1–17 (1997).
24. Bloch, R. J. & Pumplin, D. W. Molecular events in synaptogenesis: nerve-muscle adhesion and postsynaptic differentiation. *Am. J. Physiol.* **254,** C345–C364 (1988).
25. Arber, S. *et al.* Requirement for the homeobox gene *Hb9* in the consolidation of motor neuron identity. *Neuron* **23,** 659–674 (1999).
26. Thaler, J. *et al.* Active suppression of interneuron programs within developing motor neurons revealed by analysis of homeodomain factor HB9. *Neuron* **23,** 675–687 (1999).
27. Masiakowski, P. & Yancopoulos, G. D. The Wnt receptor CRD domain is also found in MuSK and related orphan receptor tyrosine kinases. *Curr. Biol.* **8,** R407 (1998).
28. Xu, Y. K. & Nusse, R. The frizzled CRD domain is conserved in diverse proteins including several receptor tyrosine kinases. *Curr. Biol.* **8,** R405–R406 (1998).
29. Hall, A. C., Lucas, F. R. & Salinas, P. C. Axonal remodeling and synaptic differentiation in the cerebellum is regulated by WNT-7a signaling. *Cell* **100,** 525–535 (2000).
30. Zhou, H., Glass, D. J., Yancopoulos, G. D. & Sanes, J. R. Distinct domains of MuSK mediate its abilities to induce and to associate with postsynaptic specializations. *J. Cell Biol.* **146,** 1133–1146 (1999).
31. Zhang, M. & McLennan, I. S. During secondary myotube formation, primary myotubes preferentially absorb new nuclei at their ends. *Dev. Dyn.* **204,** 168–177 (1995).
32. Bennett, M. R. & Pettigrew, A. G. The formation of neuromuscular synapses. *Cold Spring Harb. Symp. Quant. Biol.* **40,** 409–424 (1976).
33. Anderson, M. J. & Cohen, M. W. Nerve-induced and spontaneous redistribution of acetylcholine receptors on cultured muscle cells. *J. Physiol. (Lond.)* **268,** 757–773 (1977).
34. Frank, E. & Fischbach, G. D. Early events in neuromuscular junction formation *in vitro*. Induction of acetylcholine receptor clusters in the postsynaptic membane and morphology of newly formed nerve-muscle synapses. *J. Cell. Biol.* **83,** 143–158 (1979).
35. Hume, R. I., Role, L. W. & Fischbach, G. D. Acetylcholine release from growth cones detected with patches of acetylcholine receptor-rich membranes. *Nature* **305,** 632–637 (1983).
36. Young, S. H. & Poo, M.-M. Spontaneous release of transmitter from growth cones of embryonic neurones. *Nature* **305,** 634–637 (1983).
37. Goldman, D., Brenner, H. R. & Heinemann, S. Acetylcholine receptor α-, β-, γ-, and δ-subunit mRNA levels are regulated by muscle activity. *Neuron* **1,** 329–333 (1988).
38. Klarsfeld, A. *et al.* An acetylcholine receptor α-subunit promoter conferring preferential synaptic expression in muscle of transgenic mice. *EMBO J.* **10,** 625–632 (1991).
39. Kuromi, H. & Kidokoro, Y. Nerve disperses preexisting acetylcholine receptor clusters prior to induction of receptor accumulation in *Xenopus* muscle cultures. *Dev. Biol.* **103,** 53–61 (1984).
40. Davey, D. F. & Cohen, M. W. Localization of acetylcholine receptors and cholinesterase on nerv-contacted and noncontacted muscle cells grown in the presence of agents that block action potentials. *J. Neurosci.* **6,** 673–680 (1986).
41. Kues, W. A., Brenner, H. R., Sakmann, B. & Witzemann, V. Local neurotrophic repression of gene transcripts encoding fetal AChRs at rat neuromuscular synapses. *J. Cell Biol.* **130,** 949–957 (1995).
42. Rupp, F. *et al.* Structure and chromosomal localization of the mammalian agrin gene. *J. Neurosci.* **12,** 3535–3544 (1992).
43. Enomoto, H. *et al.* GFRα1-deficient mice have deficits in the enteric nervous system and kidneys. *Neuron* **21,** 317–324 (1998).

Acknowledgements

We thank J. Kehrl for HB9 mutant mice, and G. Yancopoulos, D. Glass and T. DeChiara for MuSK mutant mice. We thank S. Anant, C. Kintner, F. Gage and S. Heinemann for comments on the manuscript. This work was supported by grants from the NIH (J.R.S., S.L.P. and K.F.L.), the Leducq Foundation and NIH training grant (W.L.) and an NIH postdoctoral fellowship (R.W.B). S.L.P. and K.F.L are Pew Scholars.

Correspondence and requests for materials should be addressed to K.F.L. (e-mail: klee@salk.edu).

© 2001 Macmillan Magazines Ltd

A phenomenological description of space-time noise in quantum gravity

Giovanni Amelino-Camelia

Dipartimento di Fisica, Universitá "La Sapienza", P. le Moro 2, I-00185 Roma, Italy

Space-time 'foam' is a geometric picture of the smallest size scales in the Universe, which is characterized mainly by the presence of quantum uncertainties in the measurement of distances. All quantum-gravity theories should predict some kind of foam[1,2], but the description of the properties of this foam varies according to the theory, thereby providing a possible means of distinguishing between such theories. I previously showed[3] that foam-induced distance fluctuations would introduce a new source of noise to the measurements of gravity-wave interferometers, but the theories are insufficiently developed[4] to permit detailed predictions that would be of use to experimentalists. Here I propose a phenomenological approach that directly describes space-time foam, and which leads naturally to a picture of distance fluctuations that is independent of the details of the interferometer. The only unknown in the model is the length scale that sets the overall magnitude of the effect, but recent data[5] already rule out the possibility that this length scale could be identified with the 'string length' (10^{-34} m $< L_s < 10^{-33}$ m). Length scales even smaller than the 'Planck length' ($L_P \approx 10^{-35}$ m) will soon be probed experimentally.

One of the most robust[1,2] expectations for quantum gravity—the theory describing the interplay between gravity and quantum mechanics—is that space-time itself at the smallest scales should manifest quantum fluctuations of geometry. These geometrical fluctuations could then be described in terms of a highly non-trivial structure of (3+1-dimensional) space-time. This idea can be roughly visualized using an analogy with ordinary 'foamy' or 'spongy' materials, imagining however that physical processes are confined to the material of the 'sponge'. Another useful intuition-building analogy can be made with the Heisenberg uncertainty principle: whereas this principle assigns a minimum to uncertainties relevant for the combined measurement of the position and the momentum of a particle in a fixed (background) space-time, we now expect an uncertainty principle for space-time itself. This would set an absolute limit on the measurability of distances.

Attempts have been made[3,4,6–8] to describe this space-time foam through its implications for an ideal interferometer. Quantum fluctuations of distances would be observed in an interferometer as a source of noise. Theoretical predictions for this noise could be tested by comparing them with the noise levels found experimentally. In particular, a given picture of foam-induced distance fluctuations must be ruled out if it predicts more noise than the total noise found experimentally.

The first studies on this subject[3,4,6–8] indicated in various ways that the sensitivity of modern interferometers could be sufficient for the detection of space-time fluctuations originating at planckian distance scales. For experimental purposes, it would be useful to have a detailed description of the fluctuations induced by space-time foam. Unfortunately, the scarcity of experimental information on the quantum-gravity realm has prevented the development of a selection process, so there are a large number of possible quantum-gravity theories. There are two approaches whose mathematical/logical consistency has already been explored in some depth: one based on "critical superstrings"[9,10] and another based on "canonical/loop quantum gravity"[11–13]. However, even for these we do not as yet have a satisfactory understanding of their physical implications, such as the properties of space-time foam. In the few phenomenological programmes investigating other quantum properties of space-time[14–18], the difficulties deriving from the preliminary status of quantum-gravity theories have been avoided by developing direct phenomenological descriptions of the relevant phenomena. I will apply the same strategy to the description of the noise induced in interferometers by quantum gravity.

To guide interferometric studies of foam it is only necessary to estimate a relatively simple (single variable) function: the power spectrum $\rho_h(f)$ of the strain noise[19,20]. (Strain here has the standard engineering definition $h \equiv \Delta L/L$ in terms of the displacement ΔL over a given distance L.) In fact, the strain-noise power spectrum, through its dependence on the frequency f at which observations are performed, contains the most important information on the distance fluctuations, such as the mean square deviation (which is given by the integral of the power spectrum over the bandwidth of operation of the detector), and is the quantity against which the observations are compared.

The strain noise induced by quantum gravity should depend only on the Planck length, the speed-of-light constant c ($c \approx 3 \times 10^8$ m s^{-1}), and, perhaps, a length scale characterizing the properties of the apparatus with respect to quantum gravity. Within this conceptual framework there is a possible $\rho_h(f)$ estimate for foam-induced 'white noise' (noise with constant, f-independent, power spectrum). White noise is to be expected whenever the relevant stochastic phenomena are such that there is no correlation between one fluctuation and the next, an hypothesis which seems plausible for space-time fluctuations. The hypothesis that foam-induced noise is white is also consistent with the intuition emerging from analogies[21] between thermal environments and the environment provided by foam as a (dynamical) arena for physical processes. Such studies see foam-induced noise as essentially analogous to thermal noise in various physical contexts (such as electric circuits, where noise is generated by the thermal agitation of the electrons), which is indeed white whenever the bandwidth of interest is below some characteristic (resonant) frequency. In the case of foam-induced noise the characteristic frequency (which should be somewhere near the quantum-gravity frequency scale c/L_P) would be much higher than the frequencies of operation of our interferometers, and foam noise would be white at those frequencies.

Within a white-noise model, by observing that the strain-noise power spectrum has units of Hz^{-1}, I estimate that:

$$\rho_h(f) = \text{constant} \approx \frac{L_P}{c} \approx 5 \times 10^{-44}\ \text{Hz}^{-1} \qquad (1)$$

Also, because the frequencies that we can access experimentally are much smaller than c/L_P, white noise is actually the only possible structure for foam-induced strain noise if this noise is independent of the characteristics of the apparatus which is used as a space-time probe. In fact, this hypothesis implies that ρ_h can only depend on its argument f, on the Planck length and on the speed-of-light constant, and therefore the most general low-frequency expansion is of the type:

$$\rho_h(f) = a_0 \frac{L_P}{c} + a_1 \left(\frac{L_P}{c}\right)^2 f + a_2 \left(\frac{L_P}{c}\right)^3 f^2 + \ldots \qquad (2)$$

where the a_i are numerical coefficients and all monomials of the type $f^{-|n|}$ were not included in the expansion because they would require coefficients of the type $L_P^{-|n|+1}$ (which would be inconsistent with the fact that quantum-gravity effects must disappear in the limit $L_P \to 0$). For $f \ll c/L_P$ the expansion (2) is well approximated by its first term, which corresponds to the dimensional estimate (1). From the point of view of experimental tests it is also important to consider the value of the coefficient a_0, that is, to take into account the inherent uncertainty associated with the dimensional estimate (1). In this type of study, based on dimensional analysis, it is often correct to estimate that a_0 is order 1, but sometimes the dimensional

© 2001 Macmillan Magazines Ltd

estimate and the experimental result disagree by a few orders of magnitude. In testing estimate (1) we shall therefore be looking for sensitivities extending a few orders of magnitude below the L_P/c level.

In the same sense that equation (1) provides a plausible $\rho_h(f)$ estimate for foam-induced noise in quantum-gravity theories with ordinary point-like (particle) fundamental objects, in theories with extended (for example, string-like) fundamental objects characterized by a length scale L_s it seems natural to consider the low-frequency estimate:

$$\rho_h \approx \frac{L_s}{c} \tag{3}$$

In string theories L_s would be the string length, which is expected to be somewhere between a factor 10 and a factor 100 larger than the Planck length, and therefore for L_s/c there is a range of values: $5 \times 10^{-43}\,\mathrm{Hz}^{-1} < L_s/c < 5 \times 10^{-42}\,\mathrm{Hz}^{-1}$.

Because they predict no dependence on the nature of the apparatus being used to probe space-time, these estimates (1) and (3) can be tested using any detector with sensitivity to distance strain, such as interferometers and resonant-bar detectors. In spite of the smallness of the effects predicted, such experiments can reach levels of sensitivity high enough to test estimates (1) and (3) completely (either discovered or ruled out) within a few years[19].

I denote the total strain noise power spectrum observed by the experiments as ρ_h^{total}. The present level of interferometric data is best characterized by the results obtained by the 40-metre interferometer[5] at Caltech and the TAMA interferometer at the Mitaka campus of the Japanese National Astronomical Observatory. (Updated information on the progress of observations performed by the TAMA interferometer can be found at http://tamago.mtk.nao.ac.jp.) Both have obtained ρ_h^{total} values of order $10^{-40}\,\mathrm{Hz}^{-1}$ (the lowest level has been achieved by TAMA at around 1 kHz: $\rho_h^{\mathrm{total}} \approx 3 \times 10^{-41}\,\mathrm{Hz}^{-1}$). The present sensitivity of $\rho_h^{\mathrm{total}} \approx 5 \times 10^{-43}\,\mathrm{Hz}^{-1}$ of resonant-bar detectors such as NAUTILUS[22] (which is achieved near 924 Hz) is even better. This is already quite close to the natural quantum-gravity estimate L_P/c and is already at the level of L_s/c. Experiments are already probing a potentially interesting region and in order to complete a satisfactory test of estimates (1) and (3) we need only improve the sensitivity by a few orders of magnitude (in order to exclude also the possibility that the coefficient a_0 be smaller than 1).

This will be accomplished in the near future. Planned upgrades of the NAUTILUS resonant-bar detector are expected[22,23] to reach a sensitivity of $7 \times 10^{-45}\,\mathrm{Hz}^{-1}$. The LIGO/VIRGO generation of interferometers[24,25] should achieve a sensitivity or the order of $10^{-44}\,\mathrm{Hz}^{-1}$ within a year or two, during its first phase of operation. (Updated information on expected sensitivity of an advanced phase of the LIGO interferometer can be found at http://www.ligo.caltech.edu/ligo2/.) In a few years, with the space interferometer LISA[26] and especially with the "advanced phase"[23,24] of the LIGO/VIRGO interferometers, another significant sensitivity improvement should be achieved: according to recent estimates[24] it should be possible to reach sensitivity levels in the region of $10^{-48}\,\mathrm{Hz}^{-1}$, more than four orders of magnitude below the natural L_P/c estimate considered here.

This expected experimental progress is described in Fig. 1, together with the L_P/c white-noise level and the analogous noise-level predictions that can be obtained by assuming instead that the foam-induced noise be of the 'random walk' type (that is, with f^{-2} frequency dependence of the power spectrum[20]). Through the example of random-walk noise, Fig. 1 shows that the sensitivity of modern interferometers also affects non-white-noise models of foam-induced noise. This is an important consideration in assessing the overall significance of the interferometric studies considered here; in fact, the quantum-gravity realm is very far from the experimental contexts that formed our intuition, and although the simple L_P-linear white-noise model may seem natural at present, it is reassuring that this experimental programme will be able to explore a wide class of noise models.

The example of random-walk noise can also be used to illustrate the implications of noise that, unlike L_P-linear white noise, necessarily depends on some experiment-characteristic length scale Λ. A model with random-walk strain noise linearly suppressed by the Planck length would have to predict a power spectrum of the form $\rho_h \approx cL_P f^{-2}\Lambda^{-2}$. The results of experimental tests of such a model can be described with the range of values of Λ which can be excluded. As shown in Fig. 1, for the L_P-linear random-walk-noise model the excluded range of values of Λ extends all the way up to values of Λ of the order of the optical length of the arms of the interferometer. For the random-walk case, experiments will soon even reach some sensitivity to models with effects quadratically suppressed by the Planck length; in fact, as shown in Fig. 1, the LISA interferometer[26] will be able to test the possibility of noise levels of the type $\rho_h \approx cL_P^2 f^{-2}\Lambda^{-3}$ for plausible values of the experiment-characteristic length scale Λ. Other quantum-gravity-motivated

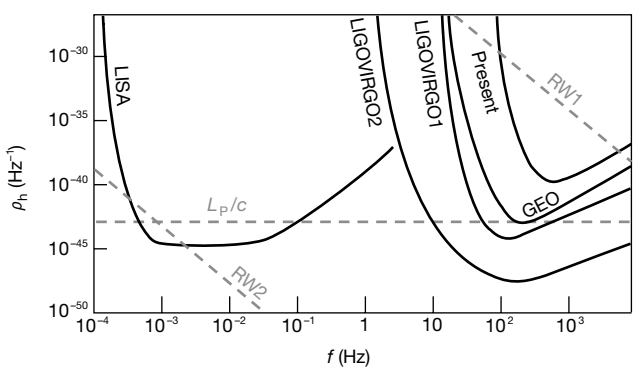


Figure 1 A qualitative (or at best semi-quantitative) comparison between the sensitivity of some interferometers and types of strain-noise power spectra. The evolution from the level of sensitivity (Present) of interferometers already in operation, to GEO (GEO) and the first phase of the LIGO and VIRGO interferometers (LIGOVIRGO1), and finally to LISA (LISA) and the second phase of LIGO and VIRGO (LIGOVIRGO2), will go through some significant phenomenological milestones among possible foam-induced noise levels. The white-noise line (L_P/c) will be crossed already by the first phase of LIGO and VIRGO. The random-walk line (RW1) has magnitude suppressed linearly by the Planck length, and is ruled out by the data of the 'Present' line. With LISA we will even start probing a small range of values of the overall coefficient c/Λ^3 (where Λ should be a scale characteristic of the experimental set-up) of the case with random-walk noise levels suppressed by the square of the Planck length. In fact, the 'RW2' line corresponds to $\rho_h \approx cL_P^2/(\lambda^3 f^2)$, where λ is the wavelength of the LISA beam.

© 2001 Macmillan Magazines Ltd

experimental programmes can only achieve sensitivity to effects that are linear in the Planck length[8,14–18]; thus LISA's capability to reach 'L_P^2 sensitivity' will mark the beginning of another significant phase in the search of quantum properties of space-time.

To make the phenomenological approach that I propose more powerful, the most urgent challenge to theory concerns the understanding of the role of energy considerations in quantum gravity. A similar phenomenological approach applied to the analysis of other interferometric noise sources would easily find ways of discriminating between different noise models by evaluating the amount of energy required by the corresponding fluctuation schemes[23]. Unfortunately, although energy considerations are rather elementary when the analysis is supported by a fixed-background space-time, the fact that quantum gravity cannot rely on a background space-time makes energy considerations much more subtle[27]. This issue will be discussed in detail elsewhere[28].

Experimentally, my analysis of quantum-gravity noise provides additional motivation for the studies planned by LISA and for the "advanced phase" of the LIGO/VIRGO interferometers. The original classical-gravity objective of modern interferometers—the discovery of Einstein's gravity waves—might well be achieved already by the "first phase" of LIGO/VIRGO, but even if so it seems to be necessary to maintain the present ambitious sensitivity objectives for LISA and the LIGO/VIRGO "advanced phase". The result could be the first experimental evidence of a quantum property of space-time. □

Received 24 May 2000; accepted 7 March 2001.

1. Wheeler, J. A. From relativity to mutability. *Revista Mexicana De Fisica* **23,** 1–57 (1974).
2. Hawking, S. W. Spacetime foam. *Nuc. Phys. B* **144,** 349–362 (1978).
3. Amelino-Camelia, G. Gravity-wave interferometers as quantum-gravity detectors. *Nature* **398,** 216–218 (1999).
4. Amelino-Camelia, G. Gravity-wave interferometers as probes of a low-energy effective quantum gravity. *Phys. Rev. D* **62,** 024015-1–024015-18 (2000).
5. Abramovici, A. *et al.* Improved sensitivity in a gravitational wave interferometer and implications for LIGO. *Phys. Lett. A* **218,** 157–163 (1996).
6. Ahluwalia, D. V. Quantum gravity: testing time for theories. *Nature* **398,** 199 (1999).
7. Ng, Y. J. & van Dam, H. Measuring the foaminess of space-time with gravity-wave interferometers. *Found. Phys.* **30,** 795–805 (2000).
8. Amelino-Camelia, G. Are we at the dawn of quantum-gravity phenomenology? *Lect. Notes Phys.* **541,** 1–49 (2000).
9. Green, M. B., Schwarz, J. H. & Witten, E. *Superstring Theory* (Cambridge Univ. Press, Cambridge, 1987).
10. Polchinski, J. *String Theory* (Cambridge Univ. Press, Cambridge, 1998).
11. Ashtekar, A. Quantum mechanics of geometry. Preprint gr-qc/9901023 at ⟨xxx.lanl.gov⟩ (1999).
12. Gaul, M. & Rovelli, C. Loop quantum gravity and the meaning of diffeomorphism invariance. *Lect. Notes Phys.* **541,** 277–324 (2000).
13. Smolin, L. The new universe around the next corner. *Phys. World* **12,** 79–84 (1999).
14. Ellis, J., Hagelin, J. S., Nanopoulos, D. V. & Srednicki, M. Search for violations of quantum mechanics. *Nucl. Phys. B* **241,** 381–405 (1984).
15. Kostelecky, V. A. & Potting, R. CPT, strings, and meson factories. *Phys. Rev. D* **51,** 3923–3935 (1995).
16. Huet, P. & Peskin, M. E. Violation of CPT and quantum mechanics in the $K_0 - \bar{K}_0$ system. *Nucl. Phys. B* **434,** 3–38 (1995).
17. Ellis, J., Lopez, J., Mavromatos, N. E., Nanopoulos, D. & the CPLEAR Collaboration. Test of CPT symmetry and quantum mechanics with experimental data from CPLEAR. *Phys. Lett. B* **364,** 239–245 (1995).
18. Amelino-Camelia, G., Ellis, J., Mavromatos, N. E., Nanopoulos, D. & Sarkar, S. Tests of quantum gravity from observations of γ-ray bursts. *Nature* **393,** 763–765 (1998).
19. Saulson, P. R. *Fundamentals of Interferometric Gravitational Wave Detectors* (World Scientific, Singapore, 1994).
20. Radeka, V. Low-noise techniques in detectors. *Annu. Rev. Nucl. Part. Sci.* **38,** 217–277 (1988).
21. Garay, L. J. Space-time foam as a quantum thermal bath. *Phys. Rev. Lett.* **80,** 2508–2511 (1998).
22. Astone, P. *et al.* Upper limit for a gravitational-wave stochastic background with the EXPLORER and NAUTILUS resonant detectors. *Phys. Lett. B* **385,** 421–424 (1996).
23. Maggiore, M. Gravitational wave experiments and early universe cosmology. *Phys. Rep.* **331,** 283–367 (2000).
24. Abramovici, A. *et al.* LIGO: The laser interferometer gravitational-wave observatory. *Science* **256,** 325–333 (1992).
25. Caron, B. *et al.* The Virgo interferometer. *Class. Quantum Grav.* **14,** 1461–1469 (1997).
26. Danzmann, K. LISA: Laser interferometer space antenna for gravitational wave measurements. *Class. Quantum Grav.* **13,** A247–A250 (1996).
27. 't Hooft, G. Quantum gravity as a dissipative deterministic system. *Class. Quantum Grav.* **16,** 3263–3279 (1999).
28. Amelino-Camelia, G. Space-time foam. Preprint gr-qc/0104005 at ⟨xxx.lanl.gov⟩ (2001).

Correspondence and requests for materials should be addressed to the author (e-mail: amelino@roma1.infn.it).

Entanglement purification for quantum communication

Jian-Wei Pan, Christoph Simon*, Časlav Brukner & Anton Zeilinger

Institut für Experimentalphysik, Universität Wien, Boltzmanngasse 5, 1090 Wien, Austria

The distribution of entangled states between distant locations will be essential for the future large-scale realization of quantum communication schemes such as quantum cryptography[1,2] and quantum teleportation[3]. Because of unavoidable noise in the quantum communication channel, the entanglement between two particles is more and more degraded the further they propagate. Entanglement purification[4–7] is thus essential to distil highly entangled states from less entangled ones. Existing general purification protocols[4–6] are based on the quantum controlled-NOT (CNOT) or similar quantum logic operations, which are very difficult to implement experimentally. Present realizations of CNOT gates are much too imperfect to be useful for long-distance quantum communication[8]. Here we present a scheme for the entanglement purification of general mixed entangled states, which achieves 50 per cent of the success probability of schemes based on the CNOT operation, but requires only simple linear optical elements. Because the perfection of such elements is very high, the local operations necessary for purification can be performed with the required precision. Our procedure is within the reach of current technology, and should significantly simplify the implementation of long-distance quantum communication.

Within the fledgling field of quantum information[9], quantum communication has recently received much experimental attention. Entangled photons have been used to experimentally demonstrate dense coding[10,11], teleportation[12–14] and quantum cryptography[15–17]. Although these schemes are realizable for moderate distances (up to a few kilometres in the case of cryptography), serious problems occur beyond this distance scale. One of the problems is the absorption of photons in the transmission channel. Because quantum states cannot be copied[18], the only way to solve this problem is by sending large numbers of photons. Nevertheless, photons still seem to be the best carriers of quantum information over long distances.

Another problem is that the quality of the entangled states generally decreases exponentially with the channel length. However, all the above protocols require that two distant parties, usually called Alice and Bob, share entangled pairs of high quality. Fortunately, various entanglement purification schemes have been suggested[4–7], by which Alice and Bob can generate a certain number of almost perfectly entangled pairs out of a larger number of less-entangled pairs using local operations and classical communication, whose precision is independent of the imperfection of the quantum channel.

To show how entanglement purification works, the scheme introduced by Bennett *et al.*[4] is illustrated in Fig. 1. The most important practical drawback of this scheme is that it requires the CNOT operation. In fact, the same is true for all known purification schemes working for general mixed entangled states. Although certain quantum logic gates have been experimentally demonstrated—for example, in ion traps[19] and high-finesse microwave cavities[20]—there is at present no implementation of CNOT gates that could realistically be used for purification in the context of long-distance quantum communication, where the probability of

* Present address: Centre for Quantum Computation, Clarendon Laboratory, University of Oxford, Parks Road, Oxford OX1 3PU, UK.

© 2001 Macmillan Magazines Ltd

errors caused by the CNOT operation must not exceed a few per cent[8]. This is far outside the range of the present implementations. Implementing purification following a recent proposal for realizing the CNOT operation with linear optics[21] is also beyond the reach of current technology. We show here how this problem can be overcome by a general purification method that does not rely on the CNOT operation.

We consider qubits implemented as the polarization states of photons. We denote the state of a horizontally polarized photon by $|H\rangle$, and the state of a vertically polarized photon by $|V\rangle$. In the usual quantum information terminology, $|H\rangle$ and $|V\rangle$ correspond to $|0\rangle$ and $|1\rangle$.

Our purification scheme is based on a simple optical element, the polarizing beam splitter (PBS). Figure 2 shows how a PBS can be used to project two photons propagating in two different spatial modes into the subspace where they have equal polarization in the basis defined by the PBS. This is achieved by selecting only those events where there is one, and only one, photon in each output mode of the PBS.

We start the explanation of our purification scheme (shown in Fig. 3) by discussing a specific example. Suppose that Alice and Bob would like to share photon pairs in the specific maximally entangled state

$$|\Phi^+\rangle_{ab} = \frac{1}{\sqrt{2}}(|H\rangle_a|H\rangle_b + |V\rangle_a|V\rangle_b) \qquad (1)$$

where the photons at Alice's and Bob's locations are denoted by a and b, respectively. Further suppose that before purification the pairs they share are all in the mixed state

$$\rho_{ab} = F|\Phi^+\rangle_{ab}\langle\Phi^+| + (1-F)|\Psi^+\rangle_{ab}\langle\Psi^+| \qquad (2)$$

where $|\Psi^+\rangle_{ab} = (1/\sqrt{2})(|H\rangle_a|V\rangle_b + |V\rangle_a|H\rangle_b)$; that is, there is an admixture of the unwanted state $|\Psi^+\rangle_{ab}$. We note that in the state $|\Phi^+\rangle_{ab}$ the two photons have equal polarization, while in the state $|\Psi^+\rangle_{ab}$ they have opposite polarization in the H/V basis. Here we assume the special form of equation (2) for simplicity only. The generality of our scheme will be proved later on.

Our proposed scheme is rather analogous to that of Bennett *et al.*[4]. We also proceed by operating on two pairs at the same time. The main difference between Fig. 3 and Fig. 1 is that now each CNOT gate is replaced by a PBS. An essential step in our purification scheme is to select those cases where there is exactly one photon in each of the four spatial output modes, which we will refer to as 'four-mode cases'. As explained above, this corresponds to a projection onto the subspace where the two photons at the same experimental station (Alice's or Bob's) have equal polarization. We note that the polarizations at the two stations do not have to be the same.

From equation (2) it follows that the original state of the two pairs can be seen as a probabilistic mixture of four pure states: with a probability of F^2, pairs 1 and 2 are in the state $|\Phi^+\rangle_{a1b1}\cdot|\Phi^+\rangle_{a2b2}$, with equal probabilities of $F(1-F)$ in the states $|\Phi^+\rangle_{a1b1}\cdot|\Psi^+\rangle_{a2b2}$ and $|\Psi^+\rangle_{a1b1}\cdot|\Phi^+\rangle_{a2b2}$, and with a probability of $(1-F)^2$ in $|\Psi^+\rangle_{a1b1}\cdot|\Psi^+\rangle_{a2b2}$.

The cross-combinations $|\Phi^+\rangle_{a1b1}\cdot|\Psi^+\rangle_{a2b2}$ and $|\Psi^+\rangle_{a1b1}\cdot|\Phi^+\rangle_{a2b2}$ never lead to four-mode cases. This is because the two entangled photons have equal polarization in the state $|\Phi^+\rangle_{ab}$ while they have opposite polarization in the state $|\Psi^+\rangle_{ab}$. Therefore, if the polarizations on Alice's side are equal, the polarizations on Bob's side must be opposite, and vice versa. Thus, by selecting only four-mode cases we can eliminate the contribution of the cross terms. This is the basic principle of our purification method.

We now consider the two remaining combinations $|\Phi^+\rangle_{a1b1}\cdot|\Phi^+\rangle_{a2b2}$ and $|\Psi^+\rangle_{a1b1}\cdot|\Psi^+\rangle_{a2b2}$. Let us first discuss the $|\Phi^+\rangle_{a1b1}\cdot|\Phi^+\rangle_{a2b2}$ case. In Fig. 3 we show that following our protocol Alice and Bob will get the state $|\Phi^+\rangle_{a3b3}$ whenever there is exactly one photon in each output mode, that is, with a probability of 50%. In the $|\Psi^+\rangle_{a1b1}\cdot|\Psi^+\rangle_{a2b2}$ case, following the same procedure Alice and Bob will project the remaining two photons a3 and b3 into the state $|\Psi^+\rangle_{a3b3}$ with a probability of 50%.

Because the probabilities for $|\Phi^+\rangle_{a1b1}\cdot|\Phi^+\rangle_{a2b2}$ and $|\Psi^+\rangle_{a1b1}\cdot|\Psi^+\rangle_{a2b2}$ are F^2 and $(1-F)^2$, respectively, after performing the purification procedure Alice and Bob will obtain the state $|\Phi^+\rangle_{a3b3}$ with a

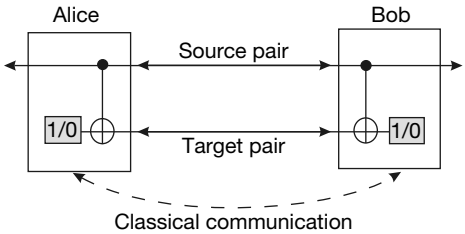


Figure 1 Scheme showing the principle of entanglement purification after Bennett *et al.*[4]. One pair of higher entanglement is created starting from two less-entangled pairs, where one member of each pair has been sent to Alice, and the other one to Bob. Both Alice and Bob perform a controlled-NOT (CNOT) operation on the two particles at their locations. Then they measure the particles belonging to the target pair in the computational basis (that is, the 0/1 basis) and compare the measured results via classical communication. If these results are the same, then the remaining pair will have a higher degree of entanglement than the original two pairs, provided that the quality of the original pairs was sufficiently high[4]. Therefore in this case they keep the source pair. In the case of obtaining opposite results, they discard it. By repeating the same procedure, always starting from the pairs produced in the former purification step, it is possible to distil pairs of arbitrarily high entanglement quality. The higher the quality desired, the more original less-entangled pairs are needed.

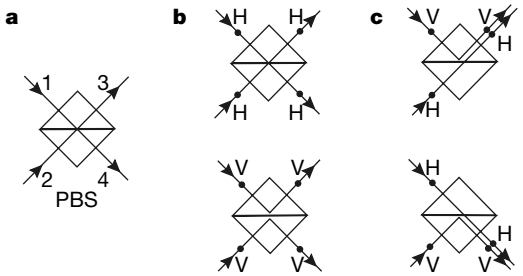


Figure 2 Using a polarizing beam splitter as a polarization comparer. **a**, The polarizing beam splitter (PBS) transmits horizontal, and reflects vertical, polarization. This means, for example, that a vertically polarized photon incident along direction 1, denoted by $|V\rangle_1$, goes out along direction 3; that is, the state $|V\rangle_1$ is transformed into $|V\rangle_3$ by the action of the PBS. Similarly, $|H\rangle_1$ goes to $|H\rangle_4$, $|H\rangle_2$ goes to $|H\rangle_3$ and $|V\rangle_2$ goes to $|V\rangle_4$. **b**, Now consider two photons incident simultaneously, one in each input mode, with equal polarization—that is, in the state $|H\rangle_1|H\rangle_2$ (top) or $|V\rangle_1|V\rangle_2$ (bottom). Then they will always go out along different directions, so there will be one photon in each of the two output modes. **c**, On the other hand, if the two incident photons have opposite polarization—that is, one is *V*- and the other is *H*-polarized—then they will always go out along the same direction, so there will be two photons in one of the two outputs and none in the other. In **b**, the two photons have been both transmitted or both reflected. This implies that finding one photon in each output mode corresponds to a projection onto the subspace spanned by $|H\rangle_1|H\rangle_2$ and $|V\rangle_1|V\rangle_2$. Ultimately, the emerging photons will not be measured in the H/V basis, but such that a coherent superposition of $|H\rangle_1|H\rangle_2$ and $|V\rangle_1|V\rangle_2$ results. This feature of the PBS, which was first described in ref. 27, has been used in the observation of multi-photon entanglement[22,23], and also in recent proposals for spin-flip-error correction[28] and entanglement concentration[29] in quantum communication. The PBS is also important in the simulation of quantum computation by linear optics[30].

© 2001 Macmillan Magazines Ltd

probability of $F^2/2$, and the state $|\Psi^+\rangle_{a3b3}$ with a probability of $(1-F)^2/2$. By applying our purification procedure, they can thus create a new ensemble described by the density operator

$$\rho'_{ab} = F'|\Phi^+\rangle_{ab}\langle\Phi^+| + (1-F')|\Psi^+\rangle_{ab}\langle\Psi^+| \qquad (3)$$

with a larger fraction $F' = F^2/[F^2 + (1-F)^2] > F$ (for $F > 1/2$) of pairs in the desired state $|\Phi^+\rangle_{ab}$ than before the purification. This concludes our discussion of purification for states of the form of equation (2). To show that our scheme also works for general input states, we now analyse in more detail the relation between our scheme and that of ref. 4.

There is a close formal correspondence between our purification scheme and the scheme of Bennett *et al.*[4] Together with the subsequent measurements, the CNOT gates in ref. 4 in fact serve the same purpose as the PBSs in our procedure: they are used by Alice and Bob to determine whether the states of the two qubits at their respective locations are equal or opposite. From the logic table of the CNOT operation ($00 \rightarrow 00$, $01 \rightarrow 01$, $10 \rightarrow 11$ and $11 \rightarrow 10$), it follows that after the operation the target particle (that is, the second) is in state 0 if originally the two particles were in equal states, and is in state 1 if they were in opposite states. In the protocol of ref. 4, after the CNOT operations on both sides Alice and Bob measure their target particles and keep the source pair if the target particles are both in state 0 or both in 1. This means that the source pair is kept in two cases: first, when the two particles on Alice's side were in equal states, and the two particles on Bob's side were also in equal states (this corresponds to the case where both targets are in 0), and second, when the two particles on Alice side were in opposite states, and the two particles on Bob's side were also in opposite states (this corresponds to the case where both targets are in 1). In these two cases, the state of the source pair will have higher entanglement than before.

The selection of four-mode cases behind the two PBS in our scheme exactly corresponds to the first case above: coincidences behind a PBS imply that the polarizations of the two incoming photons were equal. There is full formal equivalence between the two schemes, apart from one fact: in our scheme we cannot make use of the second case above. This means that for identical inputs the two procedures lead to identical outputs in the case of success, but the success probability in our case is only half as large as in the case of ref. 4. The formal equivalence of the two schemes also implies that the threshold fidelity required in order for the purification to work is the same for both, namely $F = 1/2$. Furthermore, it is now clear that general mixed input states can also be purified in our scheme by using the methods described in ref. 4, that is, by additional bilateral local operations on individual pairs. We note that our method can also be used to realize a variant of the scheme of ref. 5, which is more efficient than the scheme of ref. 4. Again the success probability is lower by a factor of 2. This is a drawback of our PBS-based scheme as it increases the number of required original entangled pairs. However, for the quantum repeater protocol[8], this increase scales only polynomially with the channel length.

We have described a full entanglement purification method that works for general mixed entangled states and does not require the CNOT operation. It only requires a parity measurement, which is non-destructive for the desired outcome. Let us now discuss the requirements for its experimental realization. Good overlap of the independently generated photons incident on each PBS is needed to erase the path information in order to make sure that the description of Fig. 2 is valid. This has been achieved in previous experiments[12,13,22,23]. An optimal realization of our scheme requires highly efficient detectors, able to distinguish between one and two photons, such that no photons are lost. Then it would be possible to infer from detections of exactly one photon each in modes a4 and b4 that there is one photon each in modes a3 and b3. However, our scheme can even be realized using detectors with limited efficiency, and which are not able to distinguish between photon numbers. The four-mode cases can be selected *a posteriori* using coincidence detection among all output modes. The inefficiency of the detectors only leads to a reduction in the overall yield, but not in the quality of the entangled pairs in modes a3 and b3. We could imagine having several subsequent purification steps. The cases where photons are missing in the output modes a3 and b3, although there was a coincidence between a4 and b4, only lead to a reduction of the efficiency of the scheme. Again, this reduction only leads to a polynomial increase in the number of pairs required for the repeater protocol. We expect that using better photon detectors, which are currently being developed[24], and fast electro-optic feed forward it will be possible to reduce these losses significantly, because then one would only proceed with the protocol when pairs are actually present.

So far, we have considered the case where it is known that at most one pair of photons is present in each of the original pairs of modes a1-b1 and a2-b2. However, the best source of entangled photons at present is spontaneous parametric down-conversion, which produces a thermal distribution of pairs, so that sometimes two pairs are emitted into the same pair of modes. This means that there are amplitudes for the photons in the four-mode case to have originated from the same source. However, by adding all relevant amplitudes coherently as described above, it is possible to see that our protocol is still successful in this situation. Although the pair production rates of current down-conversion sources are rather low, an experimental verification of our protocol with two pairs is readily achievable. It is clear that for more elaborate purification protocols involving many steps at different locations, a high-intensity source of entangled photon pairs is necessary. Substantial progress in this direction is expected: for example, improvements by several orders of magnitude were recently achieved using waveguide-enhanced down-conversion[25,26].

Protocols for true long-distance quantum communication—such as those used in quantum repeaters[8]—impose strict precision

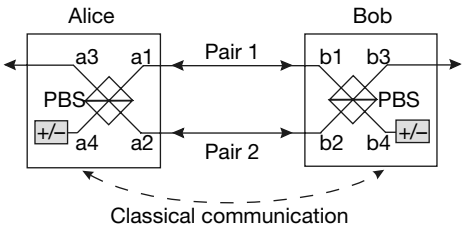


Figure 3 Our purification scheme using polarizing beam splitters. In a similar fashion to the scheme of Fig. 1, we start with two pairs shared by Alice and Bob who superimpose their photons on a PBS. They keep only those cases where there is exactly one photon in each output mode. Alice and Bob perform a polarization measurement in the 45° basis in modes a4 and b4. Depending on the results, Alice performs a specific operation on the photon in mode a3. After this procedure, the remaining pair in modes a3 and b3 will have a higher degree of entanglement than the two original pairs. To explain our protocol in detail, we consider the case where both pairs are in the state $|\Phi^+\rangle$, which occurs with a probability of F^2 in our example. For the state $|\Phi^+\rangle_{a1b1}\cdot|\Phi^+\rangle_{a2b2}$, considering only those cases for which one, and only one, photon is finally found in modes a4 and b4 results in the state $(1/2)(|H\rangle_{a3}|H\rangle_{a4}|H\rangle_{b3}|H\rangle_{b4} + |V\rangle_{a3}|V\rangle_{a4}|V\rangle_{b3}|V\rangle_{b4})$. This shows that the probability for a four-mode case is 50%. Alice and Bob can then generate maximal two-photon entanglement between the output modes a3 and b3 out of the four-photon entanglement by performing polarization measurements on each of the two photons at a4 and b4 in the +/– basis and comparing their results, where $|+\rangle = (1/\sqrt{2})(|H\rangle + |V\rangle)$ and $|-\rangle = (1/\sqrt{2})(|H\rangle - |V\rangle)$. If the measurement results at a4 and b4 are the same—that is $|+\rangle|+\rangle$ or $|-\rangle|-\rangle$—then the remaining two photons at a3 and b3 are left in the state $|\Phi^+\rangle_{a3b3}$. If the results are opposite, namely $|+\rangle|-\rangle$ or $|-\rangle|+\rangle$, then the remaining two photons are left in the state $|\Phi^-\rangle_{a3b3} = (1/\sqrt{2})(|H\rangle_{a3}|H\rangle_{b3} - |V\rangle_{a3}|V\rangle_{b3})$. In the second case, Alice can simply perform a local phase flip operation on her remaining photon to convert the state $|\Phi^-\rangle_{a3b3}$ back into $|\Phi^+\rangle_{a3b3}$.

© 2001 Macmillan Magazines Ltd

requirements on the local operations. Following our scheme, error probabilities at the one per cent level can be achieved with present technology. In particular, the precision of linear optical elements such as PBS can be extremely high, with errors as low as one-tenth of one per cent or even less. Finally, we emphasize that there is an enormous difference in the experimental effort required between implementing the CNOT operation and overlapping two photons on a PBS. We believe that the present proposal might be a vital ingredient in the future realization of long-distance quantum communication. □

Received 27 November 2000; accepted 21 February 2001.

1. Bennett, C. H. & Brassard, G. in *Proc. IEEE Int. Conf. on Computers, Systems and Signal Processing* 175–179 (IEEE, New York, 1984).
2. Ekert, A. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* **67,** 661–663 (1991).
3. Bennett, C. H. *et al.* Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70,** 1895–1898 (1993).
4. Bennett, C. H. *et al.* Purification of noisy entanglement, and faithful teleportation via noisy channels. *Phys. Rev. Lett.* **76,** 722–725 (1996).
5. Deutsch, D. *et al.* Quantum privacy amplification and the security of quantum cryptography over noisy channels. *Phys. Rev. Lett.* **77,** 2818–2821 (1996).
6. Duan, L. M., Giedke, G., Cirac, J. L. & Zoller, P. Entanglement purification of gaussian continuous variable quantum states. *Phys. Rev. Lett.* **84,** 4002–4005 (2000).
7. Bose, S., Vedral, V. & Knight, P. L. Purification via entanglement swapping and conserved entanglement. *Phys. Rev. A* **60,** 194–197 (1999).
8. Briegel, H.-J., Duer, W., Cirac, J. I. & Zoller, P. Quantum repeaters: The role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81,** 5932–5935 (1998).
9. Bennett, C. H. & DiVincenzo, D. P. Quantum information and computation. *Nature* **404,** 247–255 (2000).
10. Bennett, C. H. & Wiesner, S. J. Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Phys. Rev. Lett.* **69,** 2881–2884 (1992).
11. Mattle, K., Weinfurter, H., Kwiat, P. G. & Zeilinger, A. Dense coding in experimental quantum communication. *Phys. Rev. Lett.* **76,** 4656–4659 (1996).
12. Bouwmeester, D. *et al.* Experimental quantum teleportation. *Nature* **390,** 575–579 (1997).
13. Pan, J.-W., Bouwmeester, D., Weinfurter, H. & Zeilinger, A. Experimental entanglement swapping: Entangling photons that never interacted. *Phys. Rev. Lett.* **80,** 3891–3894 (1998).
14. Bouwmeester, D., Pan, J.-W., Weinfurter, H. & Zeilinger, A. High-fidelity teleportation of independent qubits. *J. Mod. Opt.* **47,** 279–289 (2000).
15. Jennewein, T., Simon, C., Weihs, G., Weinfurter, H. & Zeilinger, A. Quantum cryptography with entangled photons. *Phys. Rev. Lett.* **84,** 4729–4732 (2000).
16. Naik, D. S., Peterson, C. G., White, A. G., Berglund, A. J. & Kwiat, P. G. Entangled state quantum cryptography: eavesdropping on the Ekert protocol. *Phys. Rev. Lett.* **84,** 4733–4736 (2000).
17. Tittel, W., Brendel, T., Zbinden, H. & Gisin, N. Quantum cryptography using entangled photons in energy-time Bell states. *Phys. Rev. Lett.* **84,** 4737–4740 (2000).
18. Wootters, W. K. & Zurek, W. H. A single quantum cannot be cloned. *Nature* **299,** 802–803 (1982).
19. Monroe, C., Meekhof, D. M., King, B. E., Itano, W. M. & Wineland, D. J. Demonstration of a fundamental quantum logic gate. *Phys. Rev. Lett.* **75,** 4714–4717 (1995).
20. Rauschenbeutel, A. *et al.* Coherent operation of a tunable quantum phase gate in cavity QED. *Phys. Rev. Lett.* **83,** 5166–5169 (1999).
21. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409,** 46–52 (2001).
22. Bouwmeester, D., Pan, J.-W., Daniell, M., Weinfurter, H. & Zeilinger, A. Observation of three-photon Greenberger-Horne-Zeilinger entanglement. *Phys. Rev. Lett.* **82,** 1345–1349 (1999).
23. Pan, J.-W., Bouwmeester, D., Daniell, M., Weinfurter, H. & Zeilinger, A. Experimental test of quantum nonlocality in three-photon Greenberger-Horne-Zeilinger entanglement. *Nature* **403,** 515–519 (2000).
24. Takeuchi, S., Yamamoto, Y. & Hogue, H. H. Development of a high-quantum-efficiency single-photon counting system. *Appl. Phys. Lett.* **74,** 1064–1065 (1999).
25. Sanaka, K., Kawahara, K. & Kuga, T. New high-efficiency source of photon pairs for engineering quantum entanglement. Preprint quant-ph/0012028 at ⟨http://xxx.lanl.gov⟩ (2000).
26. Tanzilli, S. *et al.* Highly efficient photon-pair source using a periodically poled lithium niobate waveguide. Preprint quant-ph/0012053 at ⟨http://xxx.lanl.gov⟩ (2000).
27. Pan, J.-W. & Zeilinger, A. Greenberger-Horne-Zeilinger-state analyzer. *Phys. Rev. A* **57,** 2208–2211 (1998).
28. Bouwmeester, D. Bit-flip-error rejection in optical quantum communication. *Phys. Rev. A* **63,** R040301 (2001).
29. Yamamoto, T., Koashi, M. & Imoto, N. A concentration scheme for two partially entangled photon pairs. Preprint quant-ph/0101042 at ⟨http://xxx.lanl.gov⟩ (2001).
30. Cerf, N. J., Adami, C. & Kwiat, P. G. Optical simulation of quantum logic. *Phys. Rev. A* **57,** R1477–1480 (1998).

Acknowledgements

We thank L.-M. Duan, H. Ritsch, T. Tyc, L. Vaidman, P. Zoller and M. Zukowski for discussions. This work was supported by the Austrian Science Foundation FWF, the Austrian academy of sciences, and the TMR and QIPC programmes of the European Union.

Correspondence and requests for materials should be addressed to A.Z. (e-mail: Zeilinger-office@exp.univie.ac.at).

Mobile silver ions and glass formation in solid electrolytes

P. Boolchand & W. J. Bresser

Department of Electrical & Computer Engineering and Computer Science, University of Cincinnati, Cincinnati, Ohio 45221-0030, USA

Solid electrolytes are a class of materials in which the cationic or anionic constituents are not confined to specific lattice sites, but are essentially free to move throughout the structure. The solid electrolytes AgI and Ag_2Se (refs 1–7) are of interest for their use as additives in network glasses[8–12], such as chalcogenides and oxides, because the resulting composite glasses can show high electrical conductivities with potential applications for batteries, sensors and displays. Here we show that these composite glasses can exhibit two distinct types of molecular structures—an intrinsic phase-separation that results in a bimodal distribution of glass transition temperatures, and a microscopically homogeneous network displaying a single glass transition temperature. For the first case, the two transition temperatures correspond to the solid-electrolyte glass phase and the main glass phase (the 'base glass'), enabling us to show that the glass transition temperatures for the AgI and Ag_2Se phases are respectively 75 and 230 °C. Furthermore, we show that the magnitude of the bimodal glass transition temperatures can be quantitatively understood in terms of network connectivity, provided that the Ag^+ cations undergo fast-ion motion in the glasses. These results allow us to unambiguously distinguish base glasses in which these additives are homogeneously alloyed from those in which an intrinsic phase separation occurs, and to provide clues to understanding ion-transport behaviour in these superionic conductors.

The electrical conductivity of AgI becomes very high ($1\ \Omega^{-1}\ cm^{-1}$) at temperatures greater than 147 °C in the high-temperature phase (α-phase), when Ag^+ ions become mobile and contribute to fast-ion conduction[1]. In the companion iodide, $RbAg_4I_5$, fast-ion conduction occurs at room temperature[2], and the material has found applications as a solid electrolyte in portable batteries for heart pacemakers. The solid electrolytes AgI and Ag_2Se have also received widespread[3–7] attention as additives in network glasses, particularly chalcogenides and oxides, because of the unprecedented increase in conductivities of up to $10^{-1}\ \Omega^{-1}\ cm^{-1}$ with minuscule activation energies of 0.25 eV at room temperature. The microscopic origin of such behaviour has been the subject of continued investigations[3–7], although a consensus on the subject has yet to emerge. New insights into glass structure have recently emerged from compositional trends in glass transition temperature (T_g) that serve as a global measure of network connectivities[8]. Here we show that when the electrolytes Ag_2Se and AgI are added to chalcogenide base glasses, these electrolyte glasses in certain cases can display distinct glass transition temperatures (T_g). Constraint counting algorithms[9] reveal that the observed T_g values here provide a quantitative measure of network connectivity. The microscopic origin of glass formation is traced to Ag^+ acting as a fast diffuser, lowering the effective count of lagrangian bonding constraints[9] per atom (n_c), and promoting glass formation.

We have examined glass transitions in pseudobinary A_xB_{1-x} glasses—where A indicates additive (AgI or Ag_2Se) and B indicates chalcogenide base glass ($GeSe_4$ or As_2Se_3)—using temperature-modulated differential scanning calorimetry (MDSC). Syntheses of the glasses are described in Methods. Some of these glasses are found to be intrinsically phase-separated[10] (case I), and display bimodal T_g values—one ascribed to the base glass and the other to the additive solid-electrolyte glass phase. But some glass systems (case III) occur in which the additive and base materials

© 2001 Macmillan Magazines Ltd

homogeneously alloy to display a single T_g, which is found to decrease progressively in proportion to the additive concentration. Last, some glass systems show an intermediate behaviour (case II), in which a part of the additive phase-separates while the remainder alloys with the base glass.

Figure 1, scan B, shows an MDSC scan of $GeSe_{3.54}$ base glass containing 18.2 mol% of AgI; scan C shows As_2Se_3 base glass containing 15.0 mol% of AgI. For comparison, we have also included a heating scan of β-AgI alone (scan A); this shows the onset of the endothermic β→α transition near 147 °C. In scans B and C, there is evidence for two glass transitions. The broad endothermic feature centred near T = 75 °C, common to both glass samples, is identified as the T_g of AgI glass. The transitions at higher temperatures are identified with the base glasses[11,12] in question. AgI in $GeSe_4$ base glass (scan B) thus appears to be a model example of a system that is nearly completely phase-separated (case I). On the other hand, AgI as an additive in As_2Se_3 base glass (scan C) appears to play a dual role (case II); part of it phase-separates (displaying the T_g = 75 °C), and the rest alloys with the base glass to lower the T_g of the latter from its pristine value[12] (shown by an arrow in scan C). Furthermore, the fraction of the endotherm at T_g = 75 °C is found to grow in proportion to the molar fraction of AgI in the glasses (see Supplementary Information), supporting the suggested T_g assignments.

In Fig. 2 we show MDSC results on two base glasses, each containing 10 mol% of Ag_2Se: $GeSe_4$ (scan B) and As_2Se_3 (scan C). We also show a heating scan on β-Ag_2Se alone, showing the β→α transition (onset) near 133 °C (scan A). In scan B, there are two distinct endothermic features (case I); a primary one at T_g = 180 °C due to $GeSe_4$ base glass[11], and a secondary one at T_g = 230 °C due to Ag_2Se glass[10]; the fraction of the latter endotherm increases in proportion to the Ag_2Se molar fraction (see Supplementary Information). On the other hand, Ag_2Se added to the As_2Se_3 base glass appears to alloy homogeneously (case III); neither the T_g of Ag_2Se glass nor that of As_2Se_3 base glass (arrow in scan C) are seen, but instead there is a significantly reduced T_g (scan C). This result is not unexpected, given that a melt of equimolar concentrations of Ag_2S and As_2S_3 (that is, $AgAsS_2$) upon cooling forms a solid that can exist[13] either as a stoichiometric crystal or as a bulk glass in which Ag and As are both 3-fold coordinated, leading to a network backbone that is optimally constrained (n_c = 3).

Counting algorithms based on lagrangian bonding constraints—including bond-stretching and bond-bending—have proved to be helpful[9] in understanding the glass-forming tendency in chalcogenide glasses. Furthermore, stochastic agglomeration theory[14,15] has provided the tool to correlate quantitatively the variation of T_g with network connectivity[8]. The mechanically effective connectivity of a network can be lowered when constraints are intrinsically broken[16]. For the case of SiO_2 glass, the bond-bending constraint of oxygen atoms is intrinsically broken[16] because the bridging bond angle (Si–O–Si) in the glass network is found to display wide excursions (100–180°). These ideas can be extended to the present solid electrolytes. The structure factor of liquid Ag_2Se has been modelled[17] after α-Ag_2Se, and can also serve as a good starting point to describe the structure of the bulk glass. The rapid diffusion of Ag^+ ions through the tetrahedral interstitial sites is due to a delicate balance[18] between ionic and covalent forces; this balance makes the free energy difference, $E_6 - E_4$, between Ag taking on a coordination number of 6 (E_6) and of 4 (E_4) to nearly vanish. α-AgI possesses a fractional ionicity of 0.77 on the Phillips[18] scale, close to the threshold value of 0.785 when $E_6 = E_4$. If the I–Ag–I bond-angle flexes as Ag moves, then we expect the Ag–I–Ag bond-angle also to flex, leading to the general recognition that the bond-bending constraint about both cations (Ag^+) and anions (I^-, Se^{2-}) must be intrinsically broken in these solid electrolytes.

The structure of the superionic (α) phase of Ag_2Se and AgI is shown in Fig. 3. It consists of anions (Se, I) forming a body-centred cubic (b.c.c.) arrangement (shown blue) across which the Ag^+ cations rapidly diffuse, largely through the distorted tetrahedral interstitial sites[19] (shown red). In the case of α-Ag_2Se, on stoichiometric grounds, we must associate on average 4 Ag atoms and 2 Se atoms with the unit cell (Fig. 3), thus requiring 8 tetrahedral interstitial sites (red atoms) to be populated by Ag on the cube faces, as these are shared by two adjacent cubes. This leads to a coordination number of 4 and 8 for Ag and Se in α-Ag_2Se. For an atom possessing a coordination number, r, in a three-dimensional network, the number of independent bond-stretching constraints, n_{bs}, is given by $r/2$, while the number of bond-bending constraints, n_{bb}, is given by $2r - 3$. This gives the total number of constraints per

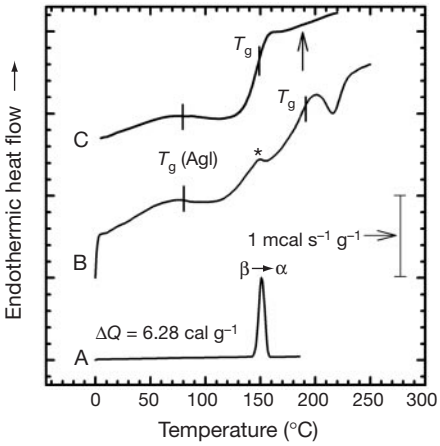


Figure 1 MDSC scans of AgI-based systems. A, the endothermic β→α transformation in β-AgI; B, two glass transitions in $Ge_{0.22}Se_{0.78}$ base glass containing 18.2 mol% AgI; C, two glass transitions in As_2Se_3 base glass containing 15 mol% AgI. In scans B and C, the broad feature with a T_g = 75 °C was due to AgI glass, while the second T_g at higher T was due to the base glasses. For the glass sample in scan B, a trace of β-AgI was quenched in, and shows evidence of the β→α transition, as indicated by the asterisk. Powder X-ray diffraction of the sample confirmed the presence of β-AgI traces. In scan C, the T_g of the base glass is shifted down by 25 °C from the T_g of pristine As_2Se_3 shown by an arrow.

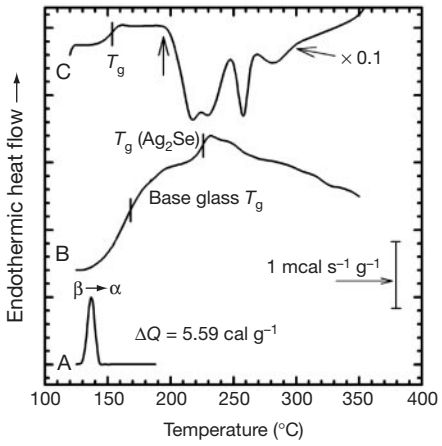


Figure 2 MDSC scans of Ag_2Se-based systems. A, β-Ag_2Se showing the β→α transition onset at 133 °C; B, $GeSe_4$ base glass containing 10 mol% Ag_2Se, showing two T_g values, one at 180 °C (ascribed to the base glass) and the other at 230 °C (ascribed to Ag_2Se glass); C, As_2Se_3 base glass containing 10 mol% Ag_2Se, showing one T_g at 160 °C ascribed to a homogeneously alloyed glass. In scan C, the reduction of the observed T_g from that in pristine As_2Se_3 (shown by arrow) is the result of the additive mixing with the base network to lower the global connectivity of the alloyed network.

© 2001 Macmillan Magazines Ltd

atom, $\bar{n}_c$, as $(5/2)\bar{r} - 3$. Enumeration of bond-stretching constraints $(r/2)$ for both Se^{2-} anions and Ag^+ cations then gives the total number of constraints for the 6 atoms in the b.c.c. cell of:

$$2[n_{bs}(Se)] + 4[n_{bs}(Ag)] = 2[4] + 4[2] = 16 \quad (1)$$

This yields $6\bar{n}_c = 16$ or $\bar{n}_c = 2.66$. The mechanically effective connectivity $(\bar{r}_m)$ of such a network is then related to $\bar{n}_c$ as follows:

$$\bar{n}_c = 5/2\bar{r}_m - 3 \quad \text{or} \quad \bar{r}_m(Ag_2Se) = 2.26 \quad (2)$$

In the case of α-AgI, on stoichiometric grounds, we must associate 2 Ag atoms with the 2 I atoms in the unit cell (Fig. 3), thus requiring 4 tetrahedral interstitial sites to be occupied by Ag. This leads to a coordination number of 4 for both Ag and I atoms. Enumeration of bond-stretching constraints for both Ag^+ and I^- gives the total number of constraints for the 4 atoms in the body-centred cubic (b.c.c.) cell:

$$2[n_{bs}(I)] + 2[n_{bs}(Ag)] = 2[2] + 2[2] = 8 \quad (3)$$

This yields $4\bar{n}_c = 8$ or $\bar{n}_c = 2$. For $\bar{n}_c = 2$, equation (2) shows the mechanically effective coordination of AgI glass to be $\bar{r}_m = 2$, suggesting that the morphological structure of the glass may be similar to that of Se glass.

In the binary chalcogenide glasses (Si-Se, Ge-Se), a linear increase of T_g values with $\bar{r}_m$ (Fig. 4) at low $\bar{r}$ (< 2.20) has been found[8]. This has been analysed[14,15] in terms of a random cross-linking of Se_n-chains by the tetrahedrally coordinated (coordination number 4) group IV (Si, Ge) atoms. The stochastic agglomeration theory[14,15] predicts a parameter-free slope, $dT_g/d\bar{r}_m$, that is in excellent accord with measured[8] T_g changes. On this plot (Fig. 4), we have also shown the observed T_g values of AgI and Ag_2Se, and find the suggestion of a sharper increase of T_g with $\bar{r}_m$ (broken curve in Fig. 4) in the present solid-electrolyte glasses. We expect the $T_g(\bar{r}_m)$ variation to exhibit a power-law behaviour, as the coordination number of the anions (I^-, Se^{2-}) increases progressively from 4 to 8 in going from AgI to Ag_2Se. (Such behaviour would be in sharp contrast to the case of the binary chalcogenides, in which increases in global connectivity are brought about by increasing the content of group IV atoms possessing a fixed coordination number of 4.) This idea could be tested by examining $T_g(z)$ variation in $(AgI)_{1-z}(Ag_2Se)_z$ alloy glasses connecting the two end-member compositions ($z = 0,1$), and analysing the results using stochastic agglomeration theory: we are at present doing this. If Ag were not a fast diffuser in the electrolyte melts, the mechanically effective coordination numbers of Ag_2Se and AgI would be rather high and would preclude[9] glass formation. Thus, constraint-counting algorithms reveal the connection between glass-forming tendency and fast-ion conduction in these archetypal solid electrolytes.

When AgI and Ag_2Se are added to chalcogenide base glasses, they form glasses largely when the base glass is optimally constrained ($\bar{n}_c = 3$). Thus, these additives will spontaneously crystallize to β-AgI and β-Ag_2Se in an underconstrained Se base glass ($\bar{n}_c = 2$) or in an overconstrained $Ge_{0.30}Se_{0.70}$ base glass ($\bar{n}_c = 3.50$), but these electrolytes will form glasses in an optimally constrained ($\bar{n}_c = 3$) base glass such as $GeSe_4$ and As_2Se_3. This is perhaps due to the absence of connectivity-induced mechanical stress when the backbone is optimally constrained[12]. Near $\bar{n}_c = 3$, the base network is known to self-organize[20] and become stress-free[12], making it easy for the additive to finely disperse and form a bulk glass, whose signature constitutes the observation of a T_g.

The chemical alloying behaviour of the solid-electrolyte additives in the present chalcogenides may have close parallels to their behaviour in oxide glasses[3,4,6,7,21,22]. In those instances (case I) where AgI glass completely phase-separates from the base glass, we expect conductivity to display a threshold behaviour corresponding to the volume percolation transition[23] of the electrolyte phase for which there may already be electrical transport evidence[4,24]—although this needs to be confirmed structurally. On the other extreme (case III), alloys composed of a 2:1 molar ratio of AgI to Ag_2SeO_4, in which the equal-sized $(SeO_4)^{2-}$ and I^- anions can intimately mix (chemical disorder) to promote glass formation, show[21,22] a Se-glass-like T_g = 50 °C characteristic of an undercoordinated r_m = 2 network, again in harmony with the present counting of lagrangian constraints. And the enhancement of electrical conductivity in such homogeneously alloyed systems may be due to a reduced activation energy upon an increase in

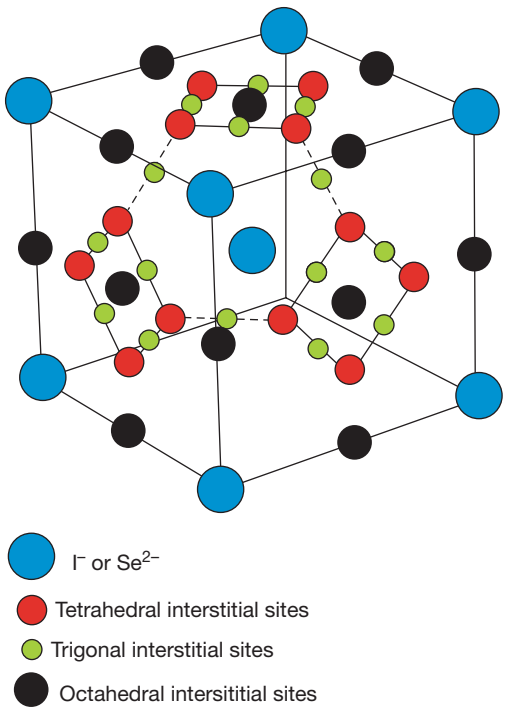


Figure 3 The crystal structure of α-Ag_2Se and α-AgI. The Se^{2-} and I^- anions (shown blue) form a b.c.c. unit, with the Ag cations rapidly diffusing across distorted tetrahedral interstitial sites (shown red) through the trigonal interstitial sites (shown green). The black circles show distorted octahedral sites.

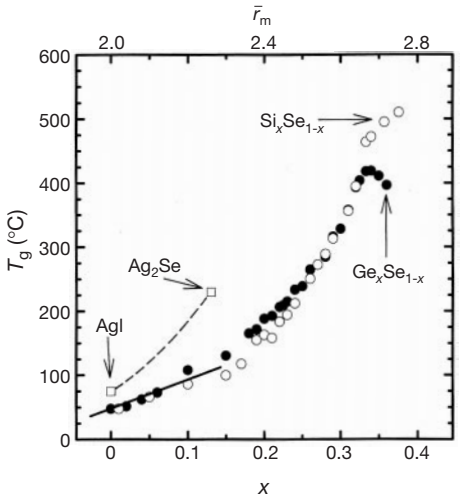


Figure 4 Variations of the glass transition temperature (T_g) of binary Ge_xSe_{1-x} and Si_xSe_{1-x} glasses. The data are taken from ref. 8, and reveal a regime ($2 < \bar{r}_m < 2.2$) of linear variation. The straight line represents the prediction of stochastic agglomeration theory[14,15]. On the other hand, T_g values in the present electrolytes may form part of a nonlinear variation sketched by the dashed line. See text for details.

© 2001 Macmillan Magazines Ltd

molar volumes[25] of the base network upon AgI addition; such increases in molar volumes also progressively lower the T_g of the alloyed glasses (Fig. 1c, Fig. 2c).

Determining the physics of transport in ionic glasses continues to be challenging. It is necessary to understand aspects of molecular structure in order to understand the details[24,26] of the conduction process. In this respect, the identification of the glass transition temperatures of AgI and Ag_2Se will also be helpful in understanding the phase separation of these additives in oxide glasses. □

Methods

Stoichiometric As_2Se_3, AgI and elemental Ge, Ag and Se (99.999% purity from Cerac, Inc.) were used as starting materials to synthesize the pseudobinary A_xB_{1-x} alloy glasses (see text). Details of glass synthesis are in our earlier[10] work, where we examined the role of Ag as an additive in Ge-Se base glasses. T_gs were measured using a model 2920 MDSC (T.A. Instruments, Inc.) at a scan rate of 3 °C min^{-1}, and modulation rate of 1 °C per 100 s.

Received 16 November 2000; accepted 20 February 2001.

1. Tubandt, C. & Lorenz, E. Molekularzustand und elektrisches Leitvermögen kristallisierter Salze. *Z. Phys. Chem.* **24,** 513–543 (1914).
2. Bradley, J. N. & Greene, P. D. Solids with high ionic conductivity in group I halide system. *Trans. Faraday Soc.* **63,** 424–430 (1967).
3. Minami, T., Imazawa, K. & Tanaka, M. Formation region and characterization of superionic conductivity glasses in the systems AgI-Ag_2O-M_xO_7. *J. Non-Cryst. Solids* **42,** 469–476 (1980).
4. Rousselot, C. *et al.* The origins of neutron scattering prepeaks and conductivity enhancement in AgI-containing glasses. *Solid State Ionics* **78,** 211–221 (1995).
5. Kincs, J. & Martin, S. W. Non-Arrhenius conductivity in glasses: mobility and conductivity saturation effects. *Phys. Rev. Lett.* **76,** 70–73 (1996).
6. Swenson, J., McGreevy, R. L., Börjesson, L., Wicks, J. D. & Howells, W. S. Intermediate-range structure of fast-ion-conductivity. *J. Phys. Condens. Matter* **8,** 3545–3552 (1996).
7. Angell, C. A. Mobile ions in amorphous solids. *Annu. Rev. Phys. Chem.* **43,** 693–717 (1992).
8. Boolchand, P. & Bresser, W. J. The structural origin of broken chemical order in $GeSe_2$ glass. *Phil. Mag. B* **80,** 1757–1772 (2000).
9. Phillips, J. C. Topology of covalent non-crystalline solids I: short range order in chalcogenide alloys. *J. Non-Cryst. Solids* **34,** 153–181 (1979).
10. Mitkova, M., Wang, Y. & Boolchand, P. Dual chemical role of Ag as an additive in chalcogenide glasses. *Phys. Rev. Lett.* **83,** 3848–3851 (1999).
11. Feng, X., Bresser, W. J. & Boolchand, P. Direct evidence for stiffness threshold in chalcogenide glasses. *Phys. Rev. Lett.* **78,** 4422–4425 (1997).
12. Georgiev, D. G., Boolchand, P. & Micoulaut, M. Rigidity transitions and molecular structure of As_xSe_{1-x} glasses. *Phys. Rev.* B **62,** R9228–R9232 (2000).
13. Slivka, V. Y., Vysochanskii, Y. M., Stefanovich, V. A., Gerasimenko, V. S. & Chepur, D. V. Features of the vibrational spectra of $AgAsS_2$. *Sov. Phys. Solid State* **24,** 392–398 (1982).
14. Kerner, R. & Micoulaut, M. On the glass transition temperature in covalent glasses. *J. Non-Cryst. Solids* **210,** 298–305 (1997).
15. Micoulaut, M. The slope equations: a universal relationship between local structure and glass transition temperature. *Eur. Phys. J. B* **1,** 277–294 (1998).
16. Zhang, M. & Boolchand, P. The central role of broken bond-bending constraints in promoting glass formation in the oxides. *Science* **266,** 1355–1357 (1994).
17. Barnes, A. C., Lague, B. S., Salmon, S. P. & Fischer, H. E. A determination of the structure of liquid Ag_2Se using neutron diffraction and isotopic substitution. *J. Phys. Condens. Matter.* **9,** 6159–6173 (1997).
18. Phillips, J. C. *Bonds and Bands in Semiconductors* 45 (Academic, New York, 1973).
19. Nield, V. M., Keen, D. A., Hayes, W. & McGreevy, R. L. Structure and fast-ion conduction in α-AgI. *Solid State Ionics* **66,** 247–258 (1993).
20. Thorpe, M. F., Jacobs, D. J., Chubynsky, M. V. & Phillips, J. C. Self-organization in network glasses. *J. Non-Cryst. Solids* **266–269,** 859–866 (2000).
21. Cramer, C., Price, D. L. & Saboungi, M. L. Structure of AgI/Ag_2SeO_4 fast-ion conducting glasses: neutron diffraction experiments. *J. Phys. Condens. Matter.* **10,** 6229–6242 (1998).
22. Shastry, M. C. R. & Rao, K. J. Physico-chemical investigation of fast ion conducting AgI-Ag_2SeO_4 glasses. *Proc. Ind. Acad. Sci.* **102,** 541–553 (1990).
23. Zallen, R. *The Physics of Amorphous Solids* 190 (Wiley & Sons, New York, 1998).
24. Bunde, A., Funke, K. & Ingram, M. D. Ionic glasses: History and challenges. *Solid State Ionics* **105,** 1–13 (1998).
25. Swenson, J. & Börjesson, L. Correlation between free volume and ionic conductivity in fast ion conducting glasses. *Phys. Rev. Lett* **7,** 3569–3572 (1996).
26. Roling, B., Happe, A., Funke, K. & Ingram, M. D. Carrier concentrations and relaxation spectroscopy: new information from scaling properties of conductivity spectra in ionically conducting glasses. *Phys. Rev. Lett.* **78,** 2160–2163 (1997).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank D.H. McDaniel, M. Mitkova, D. Georgiev, B. Goodman and M. Micoulaut for discussions. This work was supported by the Solid State Physics Program of the National Science Foundation.

Correspondence and requests for materials should be addressed to P.B. (e-mail: punit.boolchand@uc.edu).

Absence of deep-water formation in the Labrador Sea during the last interglacial period

C. Hillaire-Marcel*, A. de Vernal*, G. Bilodeau* & A. J. Weaver†

* *GEOTOP, Université du Québec à Montréal, CP 8888, Montreal, Quebec, H3C 3P8 Canada*

† *School of Earth and Ocean Sciences, University of Victoria, PO Box 3055, British Colombia, V8W 3P6, Canada*

The two main constituent water masses of the deep North Atlantic Ocean—North Atlantic Deep Water at the bottom and Labrador Sea Water at an intermediate level—are currently formed in the Nordic seas and the Labrador Sea, respectively[1]. The rate of formation of these two water masses tightly governs the strength of the global ocean circulation and the associated heat transport across the North Atlantic Ocean[2]. Numerical simulations have suggested a possible shut-down of Labrador Sea Water formation as a consequence of global warming[3]. Here we use micropalaeontological data and stable isotope measurements in both planktonic and benthic foraminifera from deep Labrador Sea cores to investigate the density structure of the water column during the last interglacial period, which was thought to be about 2 °C

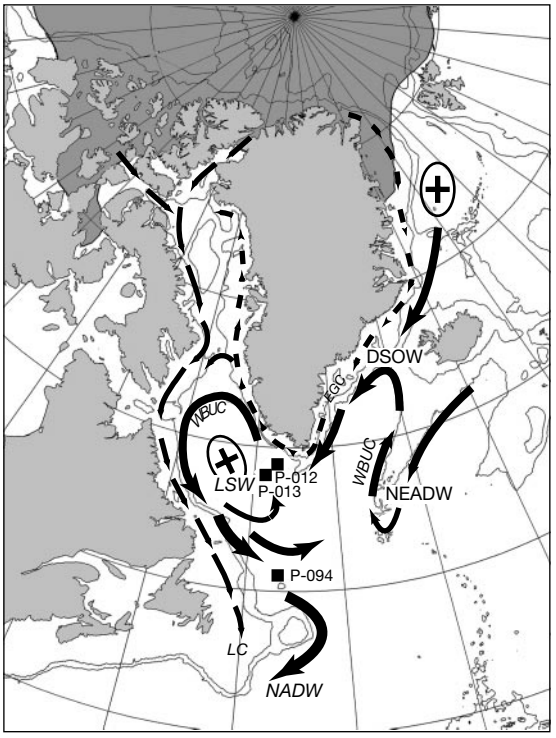


Figure 1 The Arctic–North Atlantic sector and the location of the study cores. Crosses: zones of formation of deep (DSOW) or intermediate (LSW) water. DSOW, Denmark Strait Overflow Water; LSW, Labrador Sea Water; NEADW, Northeast Atlantic Deep Water; NADW, North Atlantic Deep Water. Thin arrows, surface currents from the Fram Strait (dashed) and from the Arctic archipelago (continuous). EGC, East Greenland Current; LC, Labrador Current. Thick arrows, deep currents; WBUC, Western Boundary Undercurrent. The Arctic ice-pack (more than 11 months per year of sea-ice) is shown in dark grey. Coring sites were on the Greenland slope (P-012; 58° 55.35′ N, 47° 07.01′ W; 2,830 m), the Greenland rise (P-013; 58° 12.59′ N; 48° 22.40′ W; 3,380 m), and near Orphan knoll (P-094; 50° 12.26′ N, 45° 41.14′ W; 3,448 m).

© 2001 Macmillan Magazines Ltd

warmer than present[4]. Our results indicate that today's stratification between Labrador Sea Water and North Atlantic Deep Water never developed during the last interglacial period. Instead, a buoyant surface layer was present above a single water mass originating from the Nordic seas. Thus the present situation, with an active site of intermediate-water formation in the Labrador Sea, which settled some 7,000 years ago, has no analogue throughout the last climate cycle.

The Nordic seas are separated from the North Atlantic by the Greenland–Scotland ridge, with a maximum depth of only about 600–800 m. Hence the renewal of the deep North Atlantic water is fed by an overflow of intermediate-depth water[5]. As these waters flow southwestwards in a deep western boundary current into the Labrador Sea, they entrain recirculating, relatively cold and fresh, Labrador Sea Water (LSW)[6,7] that is largely confined to the subpolar gyre of the North Atlantic[8] (Fig. 1). The deep western boundary undercurrent leaves the Labrador Sea at depths greater than ~2,000 m; this undercurrent is thought to be about 200–300 km wide, and to transport about 13–14 Sv of newly formed North Atlantic Deep Water (NADW)[9,10] southward before eventually encountering northward-flowing Antarctic Bottom Water. Despite the high salinity of the NADW, the colder Antarctic Bottom Water has higher density and passes below the NADW (see ref. 2 for a review).

The LSW mass is renewed locally through convective mixing during winter[1] (Fig. 2), although the amount of convection varies from year to year and from decade to decade (see, for example, Fig. 2 and refs 5, 6, 8, 11). Recent simulations conducted with the Hadley Centre coupled atmosphere–ocean model have suggested that over the next few decades, LSW mass formation could be practically stopped[3] as a consequence of global warming. Attempts at reconstructing the properties of LSW formation in geological episodes warmer than the present interglacial are thus needed to develop scenarios for changes induced by global warming, and to help in evaluating the projections of coupled models. The last interglacial—isotopic substage 5e (ISS-5e)[12]—constitutes an appropriate interval, as it experienced warmer conditions than the present. Although several studies suggest a high production rate of NADW during ISS-5e (see, for example, refs 13–15), the status of the Labrador Sea as a source of intermediate or upper North Atlantic deep water is still unclear. Here we address this issue, based on a reconstruction of past gradients of potential density (σ_θ) in the water column off southwestern Greenland (Fig. 1).

The reconstruction of conditions in the surface water layer rely on transfer functions based on organic-walled dinoflagellate cysts, or dinocysts, that provide estimates of salinities and temperatures for the warmest and coldest months, August and February[16]. Based on these estimates, we calculate values of σ_θ with a degree of accuracy ($\pm$0.4; see Fig. 3a and Methods) comparable to the interannual variability based on National Oceanographic Data Center (NODC) data sets. This interannual variability is clearly illustrated at the study site by the difference between profiles from the early 1990s, when enhanced production of LSW occurred[2], and the mean NODC values[17] of the preceding decades (Fig. 2).

We then reconstruct thermohaline conditions along the seasonal pycnocline between surface and underlying water masses from oxygen isotope compositions of foraminiferal calcite ($\delta^{18}O_c$) of epipelagic (*Globigerina bulloides*) and mesopelagic (*Neogloboquadrina pachyderma* left-coiling) planktonic taxa, following approaches illustrated in refs 18–21, whereas isotopic data from benthic foraminifers (*Cibicides wuellerstorfi*) provide complementary information on bottom water. The measured $\delta^{18}O_c$-values are converted into gradients of σ_θ using calibration equations derived from August and February temperatures and salinities in surface waters, assuming that they determine the seasonal range of σ_θ values (Figs 2d, 3b and Methods).

The study site is located slightly below the high-velocity core of the Western Boundary Under Current (WBUC) (core P-013; Fig. 1). This current—which at present winnows the lower slope where sedimentation rates are very low (see core P-012; Figs 1, 4a)—is responsible for sediment focusing in the deep basin. Periods with high production rate of Denmark Strait Overflow Water (DSOW), and thus high WBUC outflow, result in enhanced sedimentation rates on the rise and reduced rates on the slope (Fig. 4a, b). The Holocene and the last interglacial, associated here to ISS-5e, fall into this category, indicating high production rates of DSOW during the last interglacial as during the Holocene, in opposition to much lower production rates (when any) throughout the glacial interval.

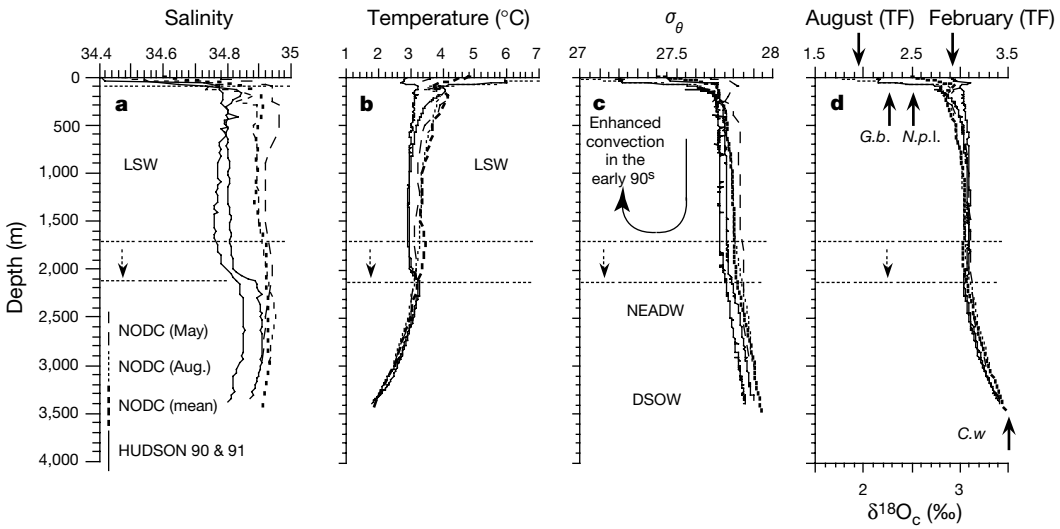


Figure 2 Structure of the modern water column at the Greenland rise site. **a**, Mean salinities from NODC data sets[17] versus 1990 and 1991 survey data on the CSS *Hudson*; **b**, mean temperatures from NODC data sets[17] versus 1990 and 1991 survey data; **c**, Potential-density values (σ_θ); **d**, Corresponding oxygen isotope composition ($\delta^{18}O_c$) for a calcite precipitated in isotopic equilibrium with ambient waters. The continuous arrows indicate $\delta^{18}O_c$ values for surface water conditions (respectively in August and February, based on transfer functions) and measured in foraminiferal assemblages from box-core sediments (*G.b.*, *Globigerina bulloides*; *N.p.*l.: *Neogloboquadrina pachyderma* left-coiling; *C.w.*, *Cibicides wuellerstorfi*). The dashed lines and arrows highlight the deepening of the pycnocline (dashed line) between the LSW and the NEADW in the early 1990s.

© 2001 Macmillan Magazines Ltd

The inception of a modern-like situation during the early Holocene (Fig. 4c) is recorded through large isotopic offsets between the benthic foraminifers bathed by the DSOW, and planktonic foraminifers which develop along the pycnocline between the LSW and the surface layer. Transcribed into potential-density gradients, and compared with surface-water estimates (Fig. 4g), these values indicate the development of a LSW/NADW stratification early during the Holocene. However, it is only after ~7,000 ^{14}C yr BP that surface-water values of σ_θ in February exceed almost permanently those recorded by *Neogloboquadrina pachyderma* left-coiling, thus suggesting winter convection and production of LSW. Similar analyses performed on a core from the outlet of the NADW into the North Atlantic, at Orphan knoll (P-094; Fig. 1), indicate the same time frame for the development of a modern-like situation[18]. The two records also suggest maximum density values in surface waters, thus maximum LSW production, during the late Holocene (Fig. 4g). Both also depict large-amplitude, millennial-scale oscillations of density in surface waters, throughout the Holocene, that probably correspond to the pervasive millennial cycle described in the North Atlantic by Bond *et al.*[22]. These authors link these oscillations to ice-rafting events. Here, they seem driven by salinity changes (Fig. 4e).

The minima in salinity and density probably correspond to enhanced freshwater supplies from the Arctic into the Labrador Sea, possibly associated with reduced sea-ice cover in the Arctic[23]. Accordingly, maxima in salinity and density, with February σ_θ values in surface waters exceeding those values recorded by planktonic foraminifers (the black arrows in Fig. 4g), should reflect reduced freshwater supplies.

The freshwater export from the Arctic to the North Atlantic Ocean is governed by the total precipitation and runoff into the Arctic. Coupled modelling studies, such as that conducted by Wood *et al.*[3], suggest that a warmer world is one in which the hydrological cycle, and hence runoff into (and precipitation over) the Arctic, is enhanced. Freshwater export from the Arctic can either be in the form of sea ice or liquid water, and can leave the Arctic into the Atlantic through either the Fram Strait or the Canadian Archipelago. The freshwater export through the Fram Strait occurs via the narrow East Greenland Current which turns into the Labrador Sea at Cape Farewell as the West Greenland Current. Analogous to conditions which occurred in the late 1960s associated with the 'great salinity anomaly'[24], we can infer that when LSW formation was not in operation, as was the case during ISS-5e (see below), enhanced freshwater export from the Arctic probably occurred. Conversely, during intervals with enhanced LSW production, the net export of fresh water from the Arctic into the Labrador Sea was reduced.

The ISS-5e section at site P-013 indeed shows features very different from those of the Holocene, notably minimum isotopic offsets between planktonic and benthic foraminifers (Figs 4c, d). Here, ISS-5e can be subdivided into two climate episodes. As shown in Fig. 4d, a maximum abundance of *Osmunda* spores at ~121 kyr ago marks the transition between them. It indicates the spreading of a fern-rich vegetation over southern Greenland and spore transport through the riverine network[25]. The *Osmunda* maximum should thus match the maximum ice retreat over southern Greenland. The interval preceding the *Osmunda* maximum shows variable but generally low surface salinity conditions, and minimum values of σ_θ (Fig. 4d, f, h), probably due to high rates of ice meltwater supplies from the retreating ice-sheet. The late ISS-5e interval corresponds to more stable conditions in surface waters, with temperatures exceeding those of the Holocene (Fig. 4e, f). Throughout both intervals, σ_θ values in surface waters remained apparently too low to permit winter convection to occur (Fig. 4h). Even at the very end of ISS-5e (at ~117 kyr ago, Fig. 4h), when maximum σ_θ values are recorded in surface waters, σ_θ values in planktonic foraminifers still remained slightly higher.

As suggested by the very narrow bracket of σ_θ values between planktonic and benthic foraminifers (Fig. 4h) a single water mass seems to have occupied the water column below the low-density surface layer throughout the entire ISS-5e episode. Therefore, the Holocene-type LSW/NADW stratification never apparently developed during ISS-5e. The water mass then occupying the Labrador Sea, from its bottom up to very near its top, had a potential density slightly greater than that of the modern LSW. It ranged from 27.80 ± 0.16 (n = 168), at the level corresponding to the habitat of

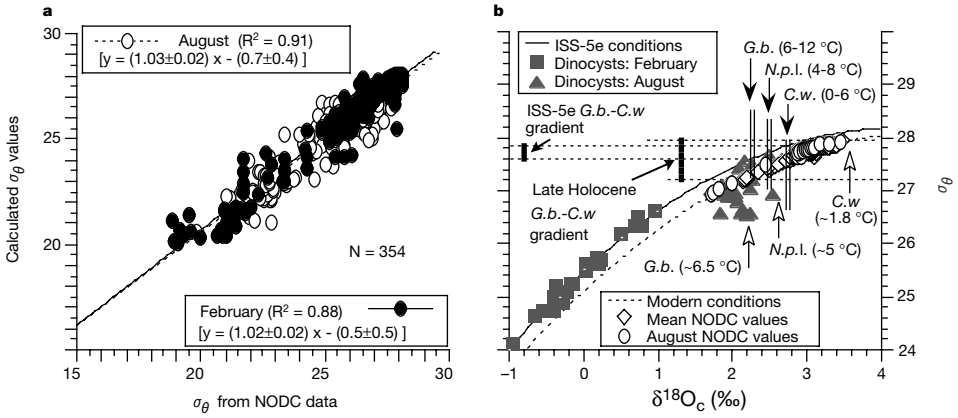


Figure 3 Estimation of potential-density values from proxies. **a**, Comparison of potential-density (σ_θ) estimates, using dinocyst assemblages from 354 core tops from the North Atlantic[16], with σ_θ values derived from NODC data sets at the corresponding sites. The strong correlation coefficients ($R^2 > 0.88$) provide an indication of the robustness of the σ_θ estimates. The scatter is due equally to variability in hydrographic measurements and to transfer-function uncertainty. The latter (that is, the standard deviation along the x axis) is $\pm 0.4\sigma_\theta$ units (±1 s.d.). **b**, Calibration of the relationship between values of σ_θ and $\delta^{18}O_c$ values in foraminiferal calcite. The equation for modern conditions is based on NODC data sets[17]: $\sigma_\theta = 25.1 + (1.3\delta^{18}O_c) - (0.14\delta^{18}O_c)^2$. The equation for ISS-5e is based on conditions in surface waters, from dinocyst transfer functions: $\sigma_\theta = 25.5 + (1.3\delta^{18}O_c) - (0.15\delta^{18}O_c)^2$. The ISS-5e equation has been constrained for high $\delta^{18}O_c$ values using maximum February σ_θ-values from transfer functions (grey triangles) and limits defined on the basis of ecological temperature ranges (the double vertical bars) of the foraminifer taxa used (the temperatures used to calculate the limits are shown in brackets). The difference between the Holocene and ISS-5e equations is essentially due to distinct boundary conditions (notably the difference of isotopic composition of fresh waters). Black arrows, mean $\delta^{18}O_c$ values of foraminiferal assemblages during ISS-5e; white arrows, mean $\delta^{18}O_c$ values during the Holocene (growth temperatures in brackets for each taxa are calculated from the projection of the measured $\delta^{18}O_c$ value on the calibration curve). Note the much narrower gradient of σ_θ values between *N.p.l.* and *C.w.* during ISS-5e compared with the Holocene.

© 2001 Macmillan Magazines Ltd

Neogloboquadrina pachyderma left-coiling, to 27.90 ± 0.12 ($n = 15$), at the sea floor (Fig. 4h), and thus compares with that of the Holocene DSOW (27.94 ± 0.02; $n = 18$; Fig. 4g). This last interglacial water mass was probably produced in the Nordic seas. A high Denmark Strait overflow rate is indeed suggested by the strong WBUC outflow of the interval (Fig. 4a, b).

Although the Holocene and ISS-5e differ in many features, both interglacials depict the same large-amplitude, millennial-scale oscillations in surface-water σ_θ values. In both cases, salinity variations seem to be the determining factor (Fig. 4e,f). Oscillations in the Arctic ice-pack and related changes in freshwater outflow routes could be invoked for both intervals. Unfortunately, direct evidence for Arctic sea-ice conditions during the last interglacial that would support this hypothesis is still missing.

We can draw three conclusions from this study. (1) No LSW formation apparently occurred during the last interglacial, or during the subsequent ice age[18], and the modern-like LSW/NADW stratification never developed during the penultimate deglaciation. LSW formation, which became established mainly after 7,000 yr BP, seems to be a feature specific to the present interglacial. (2) Both the last and Holocene interglacials depict large-amplitude, millennial-scale oscillations in surface water density, possibly linked to sea-ice conditions in the Arctic and to the subsequent routing of fresh water into the North Atlantic. (3) The maximum ice retreat over Greenland, which followed an interval of low salinity and low density related to high meltwater supply rates, was in turn followed by the establishment of relatively warm winter conditions in the Labrador Sea.

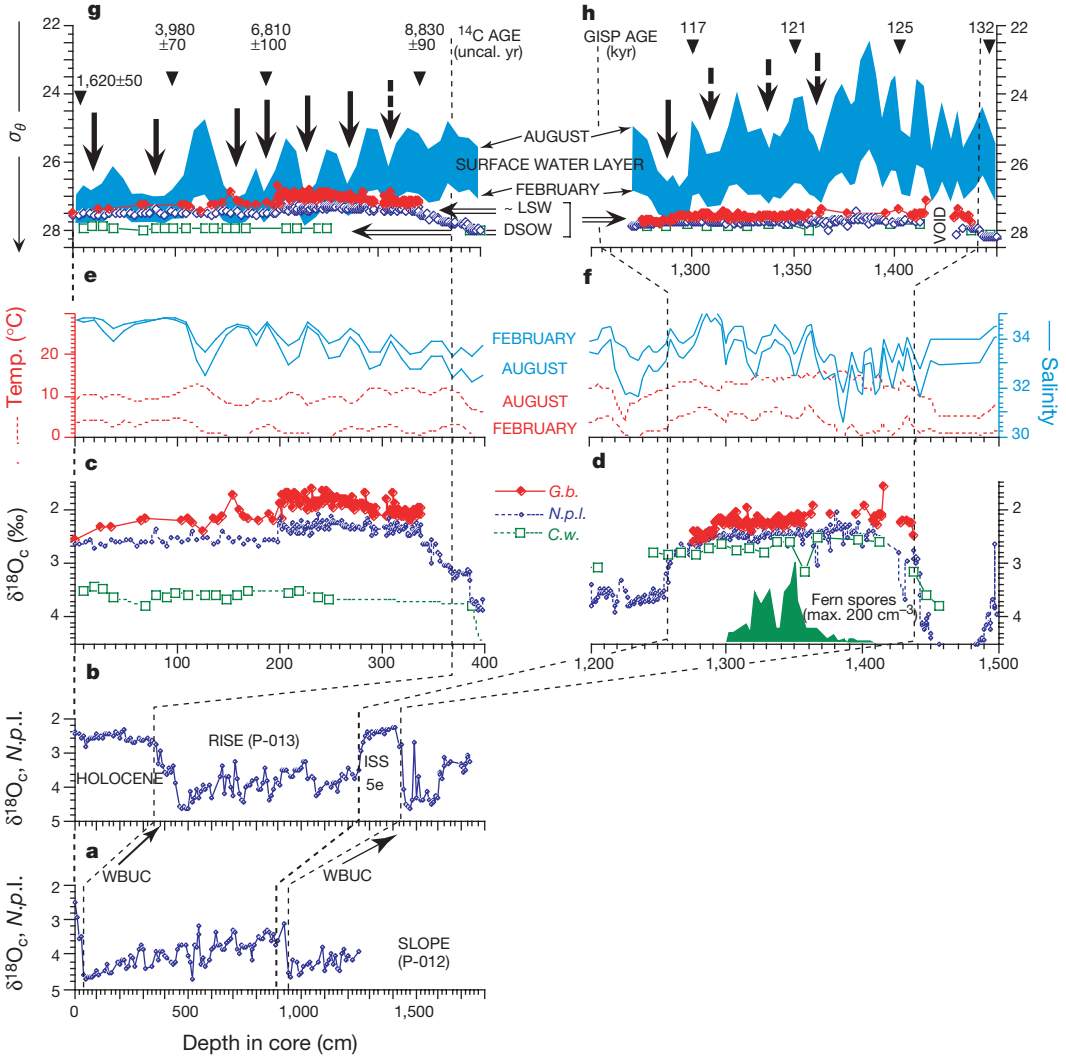


Figure 4 Magnified views of Holocene and ISS-5e records. **a** and **b**, Comparison of oxygen isotope stratigraphies at the emplacement of the high-velocity core of the WBUC (core P-012) and below it (core P-013), showing sediment focusing during interglacial intervals with high WBUC outflow (thus DSOW production)[29]. **c** and **d**, Oxygen isotope compositions ($\delta^{18}O_c$ values) of the epipelagic (*G.b.*), mesopelagic (*N.p.*l.) and benthic (*C.w.*) foraminifers during the Holocene and ISS-5e. $\delta^{18}O_c$ values in *C.w.* have been corrected by 0.64‰ for specific fractionation[26]. The fern-spore concentration curve of the ISS-5e section illustrates the maximum ice retreat over southern Greenland[25]. Complete data sets and references are available in ref. 29 and at http://www.unites.uqam.ca/geotop/ wwwgeotop.html. **e** and **f**, Temperature and salinity estimates (August versus February) from transfer functions for the Holocene and ISS-5e. **g** and **h**, Potential density gradients (reverse scale) in surface waters (grey area), from dinocyst transfer functions, and in underlying water masses, from $\delta^{18}O_c$ values in foraminifers. In **h**, the ISS-5e section is enlarged with respect to **d** and **f**. During ISS-5e, one single water mass seems to have occupied the whole water column below the dilute surface-water layer (no LSW formation, no LSW/NADW stratification). During the Holocene, notably after ~7,000 yr BP, winter σ_θ values in the surface exceed those of the underlying water mass suggested by *N.p.*l. data, indicating winter convection and LSW production. We note the large-amplitude, millennial-scale oscillations of salinity and density in surface waters during both interglacial intervals. The GISP-based chronology for the ISS-5e section is from Stoner *et al.*[30].

© 2001 Macmillan Magazines Ltd

If we view the last interglacial, thought to be about 2 °C warmer than the present[4], as an analogue for a future climate warmed by anthropogenic greenhouse gases, we might expect the eventual permanent elimination of LSW formation, in accord with the model projection of Wood *et al.*[3]. We propose that the mechanism for the eventual shutdown would involve a move towards enhanced freshwater export from the Arctic into the Labrador Sea, as a consequence of an enhanced hydrological cycle in a warmer mean climate. □

Methods

Isotopic approaches

Oxygen isotope compositions of foraminiferal calcite was determined for specimens in the 150–250 and >150 μm size fractions, respectively, for planktics and benthics, and expressed in $\delta^{18}O$ values against the PDB standard (henceforth $\delta^{18}O_c$, in ‰, with $\delta^{18}O_c = [(^{18}O/^{16}O)_{sample}/(^{18}O/^{16}O)_{standard}] - 1$). Transcription of temperatures and salinities into $\delta^{18}O_c$ values is made through the equation:

$$t = (16.9 - 4.38)(\delta^{18}O_c - A) + 0.10(\delta^{18}O_c - A)^2, \text{ with } A = (\delta^{18}O_w - 0.27)$$

where t (in °C) stands for temperature during calcite precipitation, $\delta^{18}O_c$ for the oxygen isotope composition of calcite (versus the PDB standard), and $\delta^{18}O_w$ for the oxygen isotope composition of the ambient water expressed against standard mean ocean water (SMOW)[26]. $\delta^{18}O_w$ is linked to salinity (S) through a mixing equation between a 'standard ocean' ($\delta^{18}O_{sw}$ and S_{sw}) and the local freshwater ($\delta^{18}O_{fw}$, $S \approx 0$) end-members:

$$\delta^{18}O_w \approx \delta^{18}O_{sw} - [(\delta^{18}O_{fw} - \delta^{18}O_{sw}) \times (S_w - S_{sw})/S_{sw}]$$

S_w stands for the salinity of ambient waters. $\delta^{18}O_{fw}$, $\delta^{18}O_{sw}$ and S_{sw} constitute boundary conditions that varied through time, in response to mass transfers between oceans and ice sheets, and to a lesser extent, to the meltwater release and dispersal patterns[18]. For the Holocene and ISS-5e, we used $\delta^{18}O_{sw}$ = 0‰ and S_{sw} = 35. The Holocene and modern $\delta^{18}O_{fw}$ (local) value was set to −17.8‰ (ref. 27) and the ISS-5e $\delta^{18}O_{fw}$ to −16.3‰, assuming a difference proportional to that observed in the isotopic composition of Greenland ice between the two interglacials[28].

Transfer-function estimates and the σ_θ versus $\delta^{18}O_c$ relationship

Transfer functions based on dinocysts permit the reconstruction of temperature and salinity in the photic zone (essentially, the upper mixed layer). Dinocysts are abundant in sediments from estuarine to open-ocean environments, and are thus suitable for reconstructing past conditions along continental margins of the sub-Arctic[16]. Validation exercises of transfer functions yielded the following degree of accuracy: ±1.6 °C and ±1.2 °C, for mean temperatures in August and February, respectively, and ± 0.7 salinity units in the 25–36 salinity domain[16]. Most reconstructed values fall within the range of interannual variability of the hydrographic parameters. Salinity and temperature estimates are then used to calculate σ_θ values and compared with values from NODC data[17] (Fig. 3a). The scatter of σ_θ values around the $y = x$ line has a normal distribution with a standard deviation of ±0.54 (August + February data). As interannual variations of thermohaline conditions in surface waters are almost as large, uncertainties along the x axis of Fig. 3a ($\pm\sigma_x$, linked to transfer-function reconstructions) have been considered equivalent to those along the y axis (interannual variability), allowing us to retain a σ_x value of $(0.54 \times 2^{0.5}) \approx \pm 0.40$.

The seasonal and bathymetric ranges for the development of the planktic foraminifer taxa vary depending the hydrographic settings. In the Labrador Sea, where a strong seasonal pycnocline is observed, *Globigerina bulloides* develops during warmest months only (essentially, August) towards the top of the seasonal pycnocline, in shallow and relatively low density water, whereas *Neogloboquadrina pachyderma* left-coiling develops much earlier, since May, in colder, denser and deeper waters, that is, deeper along the seasonal pycnocline (see a review in ref. 18). Accordingly, *Globigerina bulloides* shows lower $\delta^{18}O_c$ values than *Neogloboquadrina pachyderma* left-coiling, and their isotopic compositions can thus be used to estimate density gradients along the seasonal pycnocline between the surface and the sub-surface water layers[18–21]. Here, a polynomial $\delta^{18}O_c$–σ_θ relationship is calibrated from hydrographic conditions in August and February (usually the warmest and coldest months, respectively), either using NODC hydrographic data sets (modern conditions) or estimates from dinocyst transfer functions (past or modern conditions). As shown in Fig. 2d, $\delta^{18}O_c$ estimates for February and August based on transfer functions encompass isotopic compositions measured for *Globigerina bulloides* and *Neogloboquadrina pachyderma* left-coiling. The $\delta^{18}O_c$–σ_θ calibration equation, based on such estimates, thus seems suitable for interpolating σ_θ values from measurements of $\delta^{18}O_c$ in planktonic foraminifers (Fig. 3b). It has also been used for benthic foraminifers. However, because $\delta^{18}O_c$ values in benthic foraminifers fall slightly outside the range defined for surface waters (Fig. 3b), and because bottom-water masses may have different ages and hydrographic histories from surface waters, the corresponding σ_θ values are possibly less accurate.

Statistical assessment of past σ_θ values

Reconstructions of mean σ_θ values for ISS-5e and the past 7 kyr (the period of the Holocene with LWS formation) are based on 41, and 18 or more, data points, respectively. The corresponding mean σ_θ estimates for surface waters in February are of 26.72 ± 0.43 (n = 41) and 27.56 ± 0.26 (n = 18), respectively. Those calculated for the subsurface water layer, based on data from *Neogloboquadrina pachyderma* left-coiling, are 27.80 ± 0.16 (n = 168) and 27.51 ± 0.05 (n = 34), respectively. At >95% confidence interval, these values indicate density gradients compatible with winter convection, during the interval from 7 kyr ago to the present, but much too strong during ISS-5e for such deep mixing to have occurred.

Received 9 October 2000; accepted 23 February 2001.

1. Marshall, J. & Schott, F. Open-ocean convection; observations, theory, and models. *Rev. Geophys.* **37,** 1–64 (1999).
2. Weaver, A. J., Bitz, C. M., Fanning, A. M. & Holland, M. Thermohaline circulation: High latitude phenomena and the difference between the Pacific and Atlantic. *Annu. Rev. Earth Planet. Sci.* **27,** 231–285 (1999).
3. Wood, R. A., Keen, A. B., Mitchell, J. F. B. & Gregory, J. M. Changing spatial structure of the thermohaline circulation in response to atmospheric CO_2 forcing in a climate model. *Nature* **399,** 572–575 (1999).
4. White, J. W. C. Don't touch that dial. *Nature* **364,** 186 (1993)
5. Aagaard, K., Swift, J. H. & Carmack, E. C. Thermohaline circulation in the Arctic Mediterranean seas. *J. Geophys. Res.* **90,** 4833–4846 (1985).
6. Dickson, R. R. & Brown, J. The production of North Atlantic deep water: Sources, sinks and pathways. *J. Geophys. Res.* **99,** 12319–12341 (1994).
7. McCartney, M. S. Recirculating components to the deep boundary current of the northern North Atlantic. *Prog. Oceanogr.* **29,** 283–383 (1992).
8. McCartney, M. S. & Talley, L. D. Warm-to-cold water conversion in the northern North Atlantic Ocean. *J. Phys. Oceanogr.* **14,** 922–935 (1984).
9. Schmitz, W. J. Jr & McCartney, M. S. On the North Atlantic circulation. *Rev. Geophys.* **31,** 29–49 (1993).
10. Talley, L. D. & McCartney, M. S. Distribution and circulation of Labrador Sea Water. *J. Phys. Oceanogr.* **12,** 1189–1205 (1982).
11. Dickson, R. R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38,** 205–239 (1996).
12. Shackleton, N. J. Oxygen isotopes, ice volume and sea level. *Quat. Sci. Rev.* **6,** 183–190 (1987).
13. Yu, E. F., Francois, R. & Bacon, M. P. Similar rates of modern and last-glacial ocean thermohaline circulation inferred from radiochemical data. *Nature* **379,** 689–694 (1996).
14. Keigwin, L. D., Curry, W. B., Lehman, S. J. & Johnsen, S. The role of the deep ocean in North Atlantic climate change between 70 and 130 kyr ago. *Nature* **371,** 323–326 (1994).
15. Adkins, J. F., Boyle, E. A., Keigwin, L. & Cortijo, E. Variability of the North Atlantic thermohaline circulation during the last interglacial period. *Nature* **390,** 154–156 (1997).
16. de Vernal, A., Rochon, A., Turon, J.-L. & Matthiessen, J. Organic-walled dinoflagellate cysts: palynological tracers of sea-surface conditions in middle to high latitude marine environments. *Geobios* **30,** 905–920 (1997).
17. National Oceanographic Data Center *World Ocean Atlas* CR-ROM (National Oceanic and Atmospheric Administration, Boulder, 1994).
18. Hillaire-Marcel, C., de Vernal, A., Bilodeau, G. & Stoner, J. Changes of potential density gradients in the northwestern North Atlantic during the last climatic cycle based on a multiproxy approach. *Am. Geophys. Union. Monogr.* (in the press).
19. Kohfeld, K. E., Fairbanks, R. G., Smith, S. L. & Walsh, I. D. *Neogloboquadrina pachyderma* (sinistral coiling) as paleoceanographic tracers in polar oceans: Evidence from Northeast Water Polynya plankton tows, sediment traps, and surface sediment. *Paleoceanography* **11,** 679–699 (1996).
20. Andreasen, D. J & Ravelo, A. Tropical Pacific Ocean thermocline depth reconstruction for the last glacial maximum. *Paleoceanography* **2,** 395–413 (1997).
21. Faul, K., Ravelo, A. C. & Delanay, M. L. Reconstructions of upwelling, productivity, and photic zone depth in the eastern equatorial Pacific ocean using planktonic foraminiferal stable isotopes and abundances. *J. Foram. Res.* **30,** 110–125 (2000).
22. Bond, G. *et al.* Pervasive millennial-scale cycle in the North Atlantic Holocene and glacial climate. *Science* **278,** 1257–1266 (1997).
23. de Vernal, A. *et al.* Sea-ice cover and sea surface conditions in the Arctic and circum-Arctic during the Holocene: Preliminary results and methodological issues. *Can. Geophys. Union Mtg* (abstr.) (2000).
24. Dickson, R. R., Meincke, J., Malmberg, S. A. & Lee, A. J. The "great salinity anomaly" in the Northern North Atlantic 1968–1982. *Prog. Oceanogr.* **20,** 103–151 (1988).
25. de Vernal, A., Miller, G. & Hillaire-Marcel, C. Paleoenvironments of the last interglacial sensu lato in northwest North Atlantic regions: eastern Canada and adjacent seas. *Quat Int.* **10-12,** 95–106 (1991).
26. Shackleton, N. J. in *Méthodes Quantitatives d'étude des Variations du Climat au Cours du Pléistocène* (ed. Labeyrie, J.) 203–209 (Editions du CNRS, Paris, 1974).
27. Wu, G.-P. & Hillaire-Marcel, C. Oxygen isotope compositions of sinistral *Neogloboquadrina pachyderma* tests in surface sediments: North Atlantic Ocean. *Geochim. Cosmochim. Acta* **58,** 1303–1312 (1994).
28. Dansgaard, W. *et al.* Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364,** 218–220 (1993).
29. Hillaire-Marcel, C. (ed.). The Labrador Sea during the late Quaternary. *Can. J. Earth Sci. (Special issue)* **31** (1994).
30. Stoner, J. S., Channell, J. E. T. & Hillaire-Marcel, C. A 200 kyr geomagnetic stratigraphy for the Labrador Sea: Indirect correlation of the sediment record to SPECMAP. *Earth Planet. Sci. Lett.* **159,** 165–181 (1998).

Acknowledgements

This study is a contribution to the Climate System, History and Dynamics project, supported by the National Science and Engineering Research Council of Canada, and to the international Images program. Complementary support by the Fonds pour la Formation de chercheurs et l'Aide à la Recherche of Quebec Province is acknowledged.

Correspondence and requests for materials should be addressed to C.H.-M. (e-mail: chenv@uqam.ca).

© 2001 Macmillan Magazines Ltd

Evidence from the Pacific troposphere for large global sources of oxygenated organic compounds

H. Singh*, Y. Chen*, A. Staudt†, D. Jacob†, D. Blake‡, B. Heikes§ & J. Snow§

* *NASA Ames Research Center, Moffett Field, California 94035, USA*
† *Harvard University, Cambridge, Massachusetts 02138, USA*
‡ *University of California, Irvine, California 92697, USA*
§ *University of Rhode Island, Narragansett, Rhode Island 02882, USA*

The presence of oxygenated organic compounds in the troposphere strongly influences key atmospheric processes. Such oxygenated species are, for example, carriers of reactive nitrogen and are easily photolysed, producing free radicals[1–3]—and so influence the oxidizing capacity and the ozone-forming potential of the atmosphere[4–6]—and may also contribute significantly to the organic component of aerosols[7,8]. But knowledge of the distribution and sources of oxygenated organic compounds, especially in the Southern Hemisphere, is limited. Here we characterize the tropospheric composition of oxygenated organic species, using data from a recent airborne survey[9] conducted over the tropical Pacific Ocean (30° N to 30° S). Measurements of a dozen oxygenated chemicals (carbonyls, alcohols, organic nitrates, organic pernitrates and peroxides), along with several C_2–C_8 hydrocarbons, reveal that abundances of oxygenated species are extremely high, and collectively, oxygenated species are nearly five times more abundant than non-methane hydrocarbons in the Southern Hemisphere. Current atmospheric models are unable to correctly simulate these findings, suggesting that large, diffuse, and hitherto-unknown sources of oxygenated organic compounds must therefore exist. Although the origin of these sources is still unclear, we suggest that oxygenated species could be formed via the oxidation of hydrocarbons in the atmosphere, the photochemical degradation of organic matter in the oceans, and direct emissions from terrestrial vegetation.

An airborne experiment (NASA/PEM-Tropics B) that used DC-8 and P-3B aircraft provided an opportunity for a detailed examination of the composition of the oxygenated organic compounds in the remote troposphere of the Southern Hemisphere. A dozen key oxygenated chemicals and a comprehensive suite of some 28 C_2–C_8 non-methane hydrocarbons (NMHCs) were measured to an altitude of 12 km by instruments on board the DC-8 aircraft. This required the co-location and successful operation of three separate multi-channel airborne instruments (see Methods). A series of 18 DC-8 flights over the Pacific Ocean from 35° N to 35° S latitude and 90° W to 150° E longitude were performed. Nearly two-thirds of these flights were in the Southern Hemisphere. Extensive trajectory analysis and tracer measurements showed that direct impact of continental pollution in the Southern Hemisphere was minimal. Thus the air sampled here should represent near-background conditions for this season.

Figure 1 shows the mean vertical distributions of the most abundant oxygenated organic species measured in the Southern Hemisphere (Fig. 1a) and the Northern Hemisphere (Fig. 1b). The group containing reactive nitrogen is presented as total alkyl nitrates (TAN, $\Sigma RONO_2$) and peroxyacetyl nitrate (PAN, $CH_3C(O)OONO_2$). PPN (peroxypropionyl nitrate, $C_2H_5C(O)OONO_2$) was also measured, but its mixing ratios were extremely small (< 5 parts per trillion p.p.t.; 1 p.p.t. = 10^{-12} v/v) in comparison with PAN. TAN is made up of several chemicals, such as methyl nitrate, ethyl nitrate, iso-propyl nitrate and iso-butyl nitrate, which were individually measured. TAN mixing ratios in the free troposphere were ~10 p.p.t. and were as much as 60 p.p.t. in the tropical marine boundary layer, owing to the presence of a large marine source of methyl nitrate. Other alkyl nitrates can have a predominant anthropogenic source[10]. PAN has no known primary sources, and is produced as a result of atmospheric photochemical reactions between carbonyls (such as acetaldehyde and acetone) and NO_x (ref. 11). Free tropospheric (4–12 km) PAN mixing ratios in the Northern Hemisphere (60–80 p.p.t.) were substantially larger than those in the Southern Hemisphere (20–30 p.p.t.), mostly due to the greater availability of precursors. Due to its fast thermal decomposition at warm temperatures, the mixing ratio of PAN declines rapidly in the lower troposphere[11].

Mixing ratios of formaldehyde (70–300 p.p.t.) were nearly identical in the north and south tropical regions, while those of acetaldehyde and acetone were somewhat lower in the Southern Hemisphere. These acetaldehyde concentrations (60–100 p.p.t.) are the first measured (to our knowledge) for the entire remote troposphere, although limited measurements from lower troposphere are available[12]. Methanol concentrations (~900 p.p.t.) were nearly twice those of ethane, the most abundant NMHC in the troposphere. Acetone revealed a high background of about 350 p.p.t. in the Southern Hemisphere. Increased concentrations of methanol (5–20 parts per billion, p.p.b.) and acetone (2–8 p.p.b.) over rural/forested areas have been reported in recent years, and large biogenic sources have been identified[13–15]. In comparison with methanol, ethanol concentrations were rather low (< 50 p.p.t.). Methyl hydroperoxide (CH_3OOH) was ubiquitous, and one of the most abundant components in the Southern Hemisphere lower troposphere (~1,000 p.p.t.).

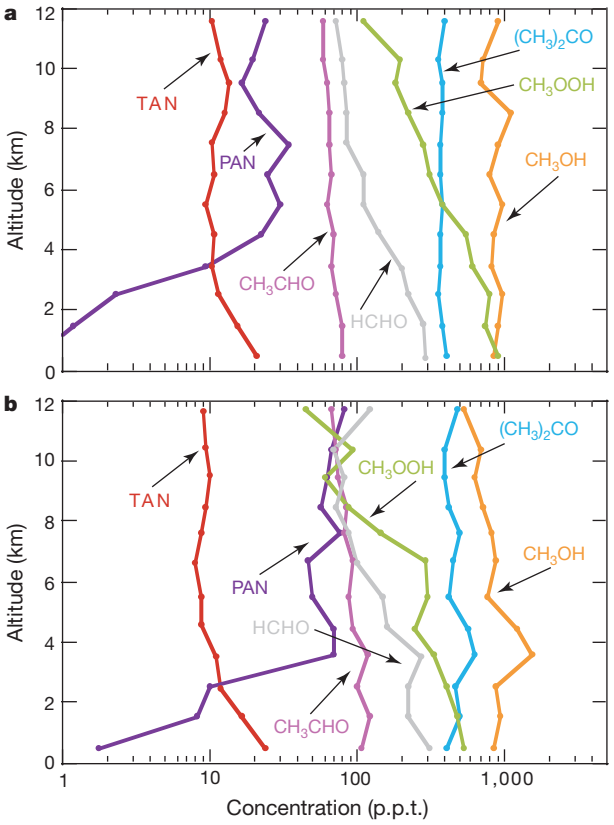


Figure 1 The mean vertical structure of oxygenated organics in the Pacific troposphere. Data were collected from the Southern Hemisphere (**a**; 0–30° S, 165° E–100° W) and the Northern Hemisphere (**b**; 0–30°N; 170–120°W). All measurements were performed over the Pacific Ocean during March–April 1999. TAN (total alkyl nitrates) represents the sum of several minor organic nitrates. PAN, peroxyacetyl nitrate.

© 2001 Macmillan Magazines Ltd

Figure 2 shows the vertical distribution of the total abundance of oxygenated organics (ΣOx-org) and C_2–C_8 NMHCs (ΣNMHCs) in the Southern Hemisphere (Fig. 2a) and the Northern Hemisphere (Fig. 2b). Although heavier hydrocarbons (> C_8) are probably present in the remote troposphere, the C_2–C_8 fraction of NMHCs should capture > 80% of the total NMHC burden[16]. At all altitudes, ΣOx-org was 2–5 times as abundant as ΣC_2–C_8 NMHCs. The mixing ratios of ΣOx-org in the Northern Hemisphere are only slightly larger than in the Southern Hemisphere (Fig. 3a), in strong contrast to NMHC (Fig. 3b). Trajectory and tracer analysis indicated that the peak in the lower troposphere near 18° N (Figs 2b and 3a) was attributable to pollution transport from the North American continent. The Northern Hemisphere sources of NMHC are well known, and the gradient of declining concentrations from the Northern to the Southern Hemisphere has been previously characterized[16]. No comparable data for the oxygenated organics have been previously available.

The mixing ratios of acetaldehyde, acetone, PAN and formaldehyde, observed over the Pacific during NASA/PEM-Tropics B, were compared with results from the Harvard global model (Fig. 4). This version of the model includes 80 chemical species (24 of them transported) to represent O_3–NO_x–hydrocarbon chemistry. It is driven by assimilated meteorological winds from the European Centre for Medium-Range Weather Forecasts, re-analysed by Florida State University to improve the quality of the assimilation in the tropics[17]. We show this comparison along flight tracks for the Southern Hemisphere (0–30° S; 165° E–100° W) and the Northern Hemisphere (0–30° N; 170–120° W). We note that the density of data from the Southern Hemisphere far exceeded that from the Northern Hemisphere. Significant disagreements between observed concentrations of acetaldehyde and acetone are present (Fig. 4a, b). The simulated acetaldehyde concentrations, which are mainly from the oxidation of ethane, are some 10 times lower than observed. Acetone mixing ratios are also lower by nearly a factor of two. PAN, a product of complex NMHC/NO_x chemistry, is underestimated in the Southern Hemisphere and overestimated in the Northern Hemisphere (Fig. 4c). Photochemically generated C_1 molecules such as formaldehyde (Fig. 4d) and CH_3OOH (not shown) were well simulated, although disagreements of the order of ± 40% were present. Methane oxidation chemistry is a major source of formaldehyde and CH_3OOH, and these two molecules can be highly effective in transferring HO_x radicals from the lower to the upper troposphere[4,6,18]. We note that the inclusion of sizeable levels of acetaldehyde and acetone in the model would cause greater PAN and formaldehyde synthesis, and some of the apparent agreements in Fig. 4 could be illusory. Lack of reliable source information on methanol prevented its simulation by the Harvard model.

We calculate that acetaldehyde has a short atmospheric lifetime of about 1 day, due to removal by reaction with the OH radical and photolysis. The measured atmospheric burden and the associated error bars can be used to estimate a large global source of 80–160 Tg yr^{-1}. Direct biogenic emissions of acetaldehyde are expected to be quite large but remain poorly quantified[19]. Acetaldehyde is known to be a common secondary product in the oxidation of many NMHCs[3] by O_3 and the OH radical. This is also the case for acetone, for which terrestrial and photochemical sources of about 60 Tg yr^{-1} have been identified[15]. Sources of acetaldehyde, acetone and methanol alone (≈ 200 Tg yr^{-1}) are estimated to be more than double the

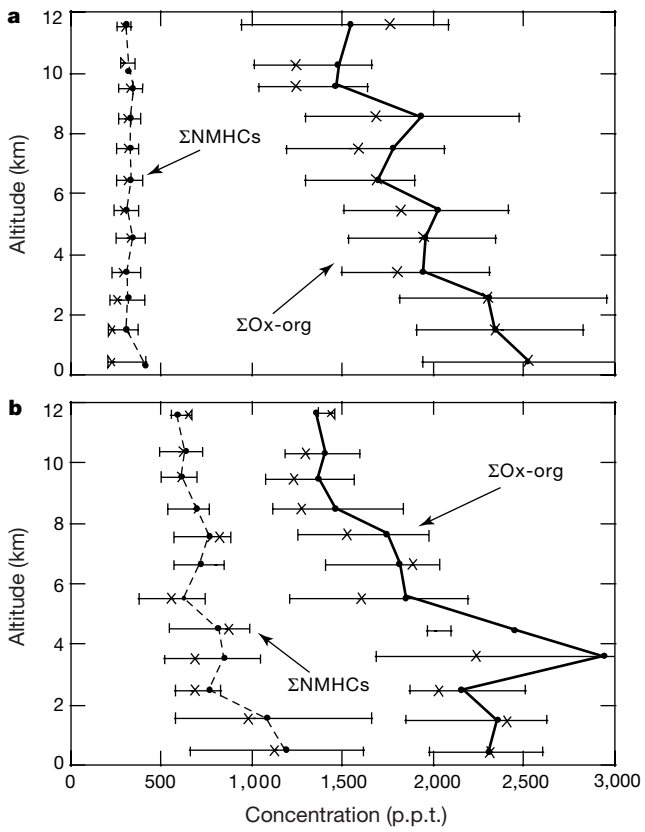


Figure 2 Distribution of total oxygenated organics (ΣOx-org) and total C_2–C_8 NMHCs (non-methane hydrocarbons; ΣNMHCs) in the remote Pacific troposphere. Data regions for **a** and **b** are as described in Fig. 1. Mean (filled circles), median (crosses), and 25 and 75 percentile (error bars) values are shown. Lines connect mean values.

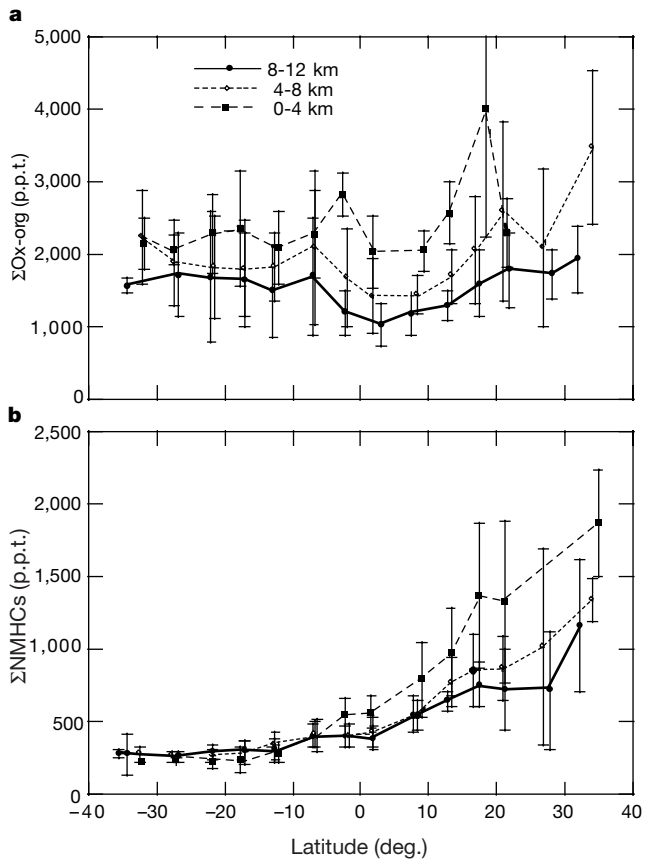


Figure 3 Latitudinal distribution of total oxygenated organics (**a**, ΣOx-org) and total C_2–C_8 NMHCs (**b**, ΣNMHCs) over the Pacific. Latitudes south of the Equator are shown as negative values. The data are divided into three altitude bands representing the lower (0–4 km), middle (4–8 km) and upper (8–12 km) troposphere. Mean values (±1σ) are shown.

© 2001 Macmillan Magazines Ltd

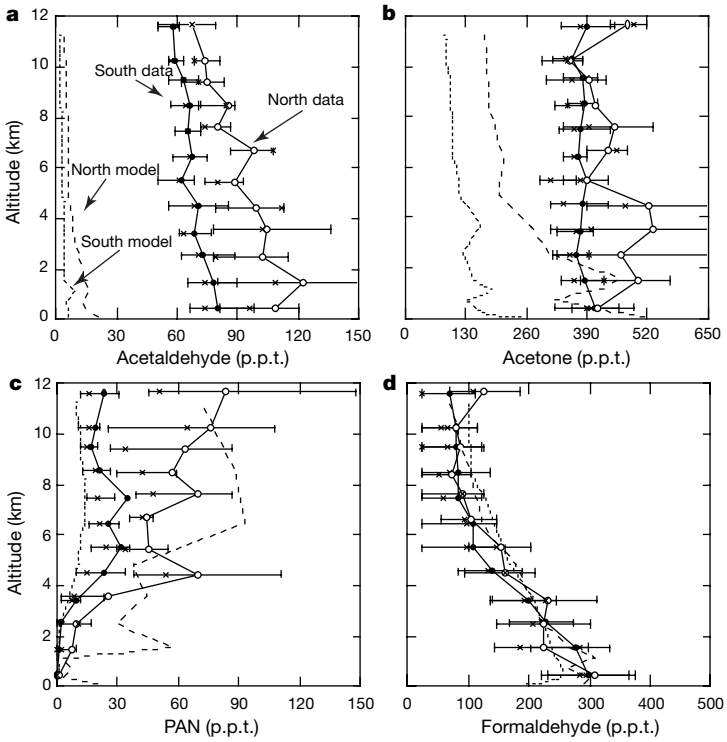


Figure 4 Comparison of measured (solid lines) and model-predicted (dashed lines) mixing ratios of four compounds in the troposphere. **a**, Acetaldehyde; **b**, acetone; **c**, PAN; and **d**, formaldehyde. A global three-dimensional photochemical model (see text) was used to calculate mixing ratios along flight tracks within the domain described in Fig. 1. Error bars as in Fig. 2.

anthropogenic emissions of NMHCs ($\sim$100 Tg yr^{-1})[15,16].

The Harvard model not only underestimates the abundance of acetaldehyde and acetone over the Pacific, it also overestimates the relative variability in the observed concentrations. These model underestimates cannot therefore be fixed by merely adding a primary source over the continents. The near-uniform observed distributions imply the presence of large diffuse sources. Such diffuse sources might arise from atmospheric oxidation of NMHCs by successive steps, each involving oxygenated intermediates with lifetime comparable to—or longer than—the parent hydrocarbon. In one model study[20], the oxidation of pentane, a highly reactive hydrocarbon, was examined in detail. Whereas nearly 97% of pentane was oxidized within a period of 5 days, 85% of the organic carbon was still present when all the complex oxygenates were included. Recently, it has been found that some 10–12% of highly reactive α/β-pinene emissions rapidly produce acetone[21]. The organic-carbon budget of the atmosphere is poorly accounted for, and this keeps open the possibility of significantly large sources of unknown oxygenates. Using an exploratory proton-transfer mass spectrometric technique, Crutzen *et al.*[22] reported detection of a variety of oxygenated organic compounds over Surinam that could not be easily explained from known tropical forest emissions. It is likely that many as-yet undiscovered oxygenated organics are present in the atmosphere.

An additional diffuse surface source for these chemicals could come from the oceans. Preliminary experiments have shown that surface sea water microlayers are highly supersaturated in chemicals such as acetaldehyde and acetone, and a group of marine bacteria that produce acetone have been isolated[23,24]. A likely source is from the photochemical degradation of dissolved organic matter present in surface oceans. Although large terrestrial sources of acetone, acetaldehyde and methanol clearly exist[15,19], the possibility that ocean biology is intricately involved in the synthesis of atmospheric oxygenated organics has not yet been carefully investigated. Lack of knowledge about the sources and fates of oxygenated organics presents an important unknown in atmospheric chemistry that needs further investigation. The production of free radicals (such as OH) by carbonyls and other oxygenates, via homogeneous and heterogeneous processes, could allow the biosphere to exert direct control over the oxidative capacity of the atmosphere. □

Methods

Details of the overall design of NASA/PEM-Tropics B and the instrumentation package used can be found in ref. 9 and references therein. Measurements reported here are based on the successful operation of three co-located multi-channel airborne instruments. These three instruments had separate inlet probes that sampled air from outside the aircraft boundary layer. Here we briefly describe these instruments and techniques.

In one three-channel instrument, a 200-ml sample of air was freeze-trapped at −140 °C and analysed by gas chromatography. PAN, PPN and alkyl nitrates were separated on two gas-chromatography columns equipped with electron-capture detectors; acetone, methanol, ethanol and acetaldehyde were measured on the third column in which a photo-ionization detector (PID) and a reduction gas detector (RGD) were placed in series. In the PID/RGD channel, the air sample was dried to −50 °C before cryogenic trapping at −140 °C. Standard addition procedures were used for in-flight calibrations. Standard mixtures in air (or N_2) were obtained using diffusion tubes for PAN, permeation tubes for carbonyls and alcohols, and pressurized cylinders for alkyl nitrates[1]. The sensitivity of detection of reactive nitrogen species was $\sim$1 p.p.t., while that of other oxygenates was $\sim$20 p.p.t.. Overall measurement accuracy for these species is estimated to be ± 25%, except for acetaldehyde, which is uncertain by a factor of two. The larger uncertainty in acetaldehyde is due to the small signal-to-noise ratio (3–5), and the difficulty in generating stable primary standards.

In a second instrument, peroxides (H_2O_2 and CH_3OOH) and formaldehyde (CH_2O) were collected in an aqueous solution. The peroxides were separated, and then quantified using a high-performance liquid chromatography fluorescence derivative technique to a sensitivity of 15–25 p.p.t. (ref. 25). Formaldehyde was quantified using a flow-injection fluorescence derivative technique to a sensitivity of 50 p.p.t. (ref. 26). Primary calibrations were based on aqueous standards in combination with state variables and measured liquid and gas mass flow rates. Collection efficiency for CH_3OOH and CH_2O was determined to vary from 50 to 75% on the basis of a series of laboratory and field tests. Standard addition procedures, in which known amounts of peroxide and CH_2O are added to the ambient air, were performed in flight.

© 2001 Macmillan Magazines Ltd

The third instrument collected whole air samples in evacuated 2-l electropolished stainless-steel canisters. The canisters were pressurized to 40 p.s.i., and then shortly after each flight were transported to a central laboratory. A 1,520-ml air sample was cryogenically trapped on a loop filled with glass beads maintained at liquid nitrogen temperature. Multi-column gas chromatography coupled with flame ionization detection and mass spectrometry was used to quantify C_2–C_8 NMHCs (alkanes, alkynes, alkenes and aromatics). Sample analysis usually occurred within one week of collection and never more than two weeks. The overall accuracy of these NMHC measurements is estimated to be 5–10%. The limit of detection was 3 p.p.t. for all NMHCs[27]. Additional details may be obtained from the authors.

Received 22 August 2000; accepted 20 February 2001.

1. Singh, H. B., Kanakidou, M., Crutzen, P. & Jacob, D. High concentrations and photochemical fate of carbonyls and alcohols in the global troposphere. *Nature* **378,** 50–54 (1995).
2. Wennberg, P. O. *et al.* Hydrogen radicals, nitrogen radicals, and the production of ozone in the upper troposphere. *Science* **279,** 49–53 (1998).
3. Carlier, P., Hannachi, H. & Mouvier, G. The chemistry of carbonyls in the atmosphere—a review. *Atmos. Environ.* **20,** 2079–2099 (1986).
4. Prather, M. & Jacob, D. A persistence imbalance in HO_x and NO_x photochemistry of the upper troposphere driven by deep tropical convection. *Geophys. Res. Lett.* **24,** 3189–3192 (1997).
5. Müller, J. F. & Brasseur, G. Sources of upper tropospheric HO_x: A three-dimensional model study. *J. Geophys. Res.* **104,** 1705–1715 (1999).
6. Jaeglé, L., Jacob, D., Brune, W. & Wennberg, P. Chemistry of HO_x radicals in the upper troposphere. *Atmos. Environ.* **35,** 469–489 (2001).
7. Murphy, D. M., Thomson, D. S. & Mahoney, M. J. *In situ* measurements of organics, meteoritic material, mercury, and other elements in aerosols at 5 to 19 kilometers. *Science* **282,** 1664–1669 (1998).
8. Li, P. *et al.* Heterogeneous reactions of VOCs on oxide particles of the most abundant crustal elements: surface reactions of acetaldehyde, acetone and propionaldehyde on SiO_2, Fe_2O_3, TiO_2 and CaO. *J. Geophys. Res.* (in the press).
9. Raper, J. L. *et al.* Pacific Exploratory Mission in the tropical pacific: PEM-Tropics B, March-April 1999. *J. Geophys. Res.* (in the press).
10. Atlas, E. *et al.* Alkyl nitrate and selected halocarbon measurements at Mauna Loa observatory, Hawaii. *J. Geophys. Res.* **97,** 10331–10348 (1992).
11. Singh, H. B. & Hanst, P. L. Peroxyacetyl nitrate (PAN) in the unpolluted atmosphere: An important reservoir for nitrogen oxides. *Geophys. Res. Lett.* **8,** 941–944 (1981).
12. Arlander, D. W., Brunking, D., Schmidt, U. & Ehhalt, D. H. The distribution of acetaldehyde in the lower troposphere during TROPOZ II. *J. Atmos. Chem.* **22,** 243–249 (1995).
13. Riemer, D. *et al.* Observations of nonmethane hydrocarbons and oxygenated volatile organic compounds at a rural site in the southeastern United States. *J. Geophys. Res.* **103,** 28111–28128 (1998).
14. Lamanna, M. S. & Goldstein, A. H. *In situ* measurements of C_2–C_{10} VOCs above a Sierra-Nevada pine plantation (including oxygenated species) from pasture. *J. Geophys. Res.* **104,** 21247–21262 (1999).
15. Singh, H. B. *et al.* Distribution and fate of selected oxygenated organic species in the troposphere and lower stratosphere over the Atlantic. *J. Geophys. Res.* **105,** 3795–3805 (2000).
16. Singh, H. B. & Zimmerman, P. B. in *Advances in Environmental Science and Technology* Vol. 24 (ed. Nariagu, J.) 177–235 (Wiley and Sons, New York, 1992).
17. Staudt, A. C. *et al.* Continental sources, transoceanic transport, and interhemispheric exchange of carbon monoxide over the Pacific. *J. Geophys. Res.* (submitted); preprint at ⟨http://www-as.harvard.edu/chemistry/trop⟩ (2001).
18. Logan, J. *et al.* Tropospheric chemistry: A global perspective. *J. Geophys. Res.* **86,** 7210–7254 (1981).
19. Kesselmeier, J. & Staudt, M. Biogenic volatile organic compounds (VOC): An overview on emissions, physiology and ecology. *J. Atmos. Chem* **33,** 23–38 (1999).
20. Madronich, S. & Calvert, J. G. Permutation reactions of organic peroxy radicals in the troposphere. *J. Geophys. Res.* **95,** 5697–5715 (1990).
21. Reissell, A. *et al.* Formation of acetone from the OH radical- and O3-initiated reactions of a series of monoterpenes. *J. Geophys. Res.* **104,** 13869–13879 (1999).
22. Crutzen, P. J. *et al.* High spatial and temporal resolution measurements of primary organics and their oxidation products over the tropical forests of Surinam. *Atmos. Environ.* **34,** 1161–1165 (2000).
23. Nemecek-Marshall, M. *et al.* Marine *Vibrio* species produce the volatile organic compound acetone. *Appl. Environ. Microbiol.* **61,** 44–47 (1995).
24. Zhou, X. & Mopper, K. Photochemical production of low-molecular-weight carbonyl compounds in seawater and surface microlayer and their air-sea exchange. *Mar. Chem.* **56,** 201–213 (1997).
25. Lee, M., Noone, B. C., O'Sullivan, D. & Heikes, B. G. Method for the collection and HPLC analysis of hydrogen peroxide and C_1 and C_2 hydroperoxides in the atmosphere. *J. Atmos. Ocean Technol.* **12,** 1060–1070 (1995).
26. Lazrus, A. L., Fong, K. L. & Lind, J. A. Automated fluorometric determination of formaldehyde in air. *Anal. Chem.* **60,** 1074–1078 (1988).
27. Blake, N. J. *et al.* Influence of southern hemispheric biomass burning on mid-tropospheric distributions of nonmethane hydrocarbons and selected halocarbons over the remote South Pacific. *J. Geophys. Res.* **104,** 16213–16232 (1999).

Acknowledgements

This research was supported by the NASA Global Tropospheric Experiment. We thank all participants in NASA/PEM-Tropics B for their support.

Correspondence and requests for materials should be addressed to H.S. (e-mail: hsingh@mail.arc.nasa.gov).

Hemispherical variations in seismic velocity at the top of the Earth's inner core

Fenglin Niu* & Lianxing Wen†

* *Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road NW, Washington DC 20015, USA*
† *Department of Geosciences, State University of New York at Stony Brook, Stony Brook, New York 11794, USA*

Knowledge of the seismic velocity structure at the top of the Earth's inner core is important for deciphering the physical processes responsible for inner-core growth[1–3]. Previous global seismic studies[4–9] have focused on structures found 100 km or deeper within the inner core, with results for the uppermost 100 km available for only isolated regions[10–12]. Here we present constraints on seismic velocity variations just beneath the inner-core boundary, determined from the difference in travel time between waves reflected at the inner-core boundary and those transmitted through the inner core. We found that these travel-time residuals—observed on both global seismograph stations and several regional seismic networks—are systematically larger, by about 0.8 s, for waves that sample the 'eastern hemisphere' of the inner core (40° E to 180° E) compared to those that sample the 'western hemisphere' (180° W to 40° E). These residuals show no correlation with the angle at which the waves traverse the inner core; this indicates that seismic anisotropy is not strong in this region and that the isotropic seismic velocity of the eastern hemisphere is about 0.8% higher than that of the western hemisphere.

PKiKP is the P wave that reflects off the inner-core boundary (ICB) and PKIKP is the P wave that propagates through the inner core. The differential travel time between these two phases is most sensitive to the seismic structure in the inner core, as these two phases have almost identical ray paths in the mantle at the distance range of 130°–142° (Fig. 1 inset). At this distance range, PKIKP samples the top 100 km of the inner core, and both PKIKP and PKiKP are identifiable in short-period seismograms. We select a total of 203 high-quality PKiKP and PKIKP waveforms from recordings in the IRIS global seismic network (GSN) and many regional seismic arrays. We choose recordings for events from 1990 to 1999 with simple source–time functions, so only those of intermediate and deep earthquakes are used. Broadband seismograms are filtered using the WWSSN short-period instrument response.

Differential travel-time residuals of PKiKP – PKIKP show a clear difference between the 'eastern hemisphere' and the 'western hemisphere' of the inner core (Fig. 1). A positive (negative) travel-time residual indicates a relatively higher (lower) seismic velocity in the top of the inner core compared to the PREM model[13]. The travel-time residuals are systematically larger in the eastern hemisphere (solid symbols) than those in the western hemisphere (open symbols), regardless of the turning depths of the PKIKP phases (Fig. 2a). These travel-time residuals, on the other hand, show no correlation with the PKIKP ray angles from the equatorial plane (Fig. 2b), indicating that anisotropy is non-existent or very weak in the outermost 100 km of the inner core. This is consistent with the travel-time and waveform studies using regional data[10–12]. On average, the residuals in the eastern hemisphere are approximately 0.8 s larger than those of the western hemisphere. The travel-time residuals observed in the eastern hemisphere can be fitted by a model with P velocities that are roughly 0.5% faster than the PREM in the top 100 km of the inner core, whereas those observed in the

© 2001 Macmillan Magazines Ltd

western hemisphere can be explained by a model with P velocities that are 0.3% slower than the PREM (Fig. 2a). Further precise models using waveform data are presented in elsewhere[14].

We only used those seismograms that have unambiguous PKiKP and PKIKP phases and clear move-outs (arrival times as a function of epicentral distance) of these two phases. The move-outs of these two phases are clear from the compilations of all observed waveforms in both hemispheres. We show an example of these move-outs of the PKiKP and PKIKP phases from record sections observed in two regional seismic arrays (Fig. 3a and b). In practice, because the onsets in the worldwide standardized seismograph network (WWSSN) responses are sometimes difficult to identify, we determine the differential travel time between PKiKP and PKIKP phases by measuring the relative timing between their maximal amplitudes. The above picking method has been shown to be very accurate in synthetic seismograms at the distance range between 130° and 141° (Fig. 3c). These synthetics are calculated by the generalized ray theory[15] using the PREM model and a source depth of 600 km. With respect to the maximal amplitudes of the PKiKP phases, the hand-picked PKIKP times based on their maximal amplitudes (vertical lines) and those calculated (dots) are in good agreement. The difference is less than 0.06 s. This difference is

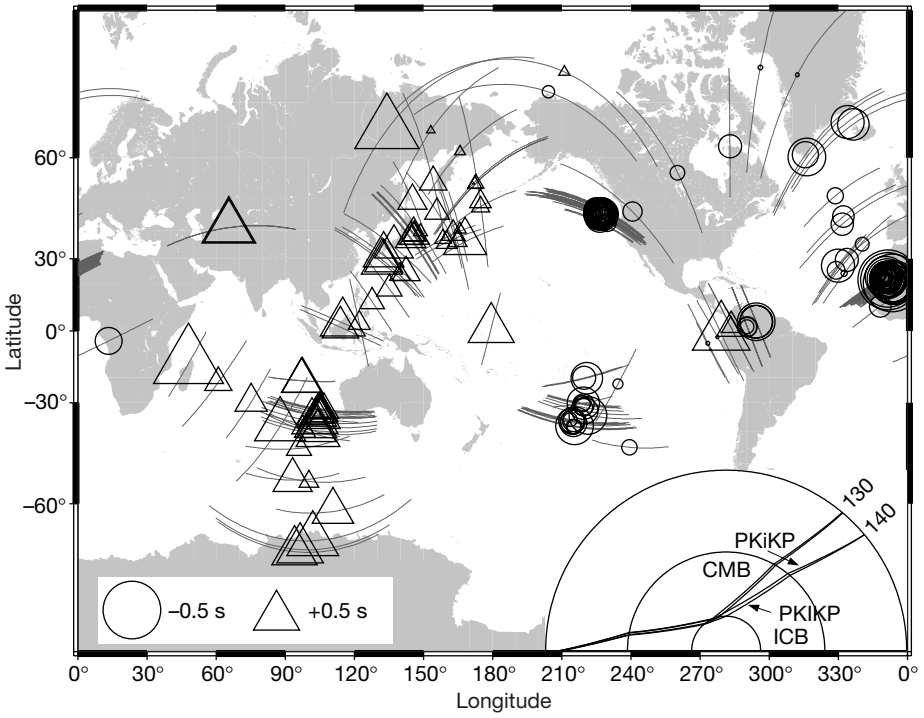


Figure 1 Map view of PKiKP minus PKIKP travel-time residuals displayed as lines along ray segments through the inner core and symbols at the turning points. The residuals are calculated with respect to the PREM. Circles and triangles represent negative and positive residuals, respectively. The size of the symbols is proportional to the absolute value of residuals. The data set was collected from the global seismic network (GSN) and many regional seismic networks: (1) the J-array[17]; (2) the Canadian national seismograph network (CNSN); (3) the Tanzania broadband seismic experiment; (4) The broadband Andean joint experiment (BANJO); (5) the BLSP 94; (6) the Tibetan Plateau passive-source seismic experiment (TIPLT). Ray paths of PKiKP and PKIKP at distances 130° and 140° are shown in the inset.

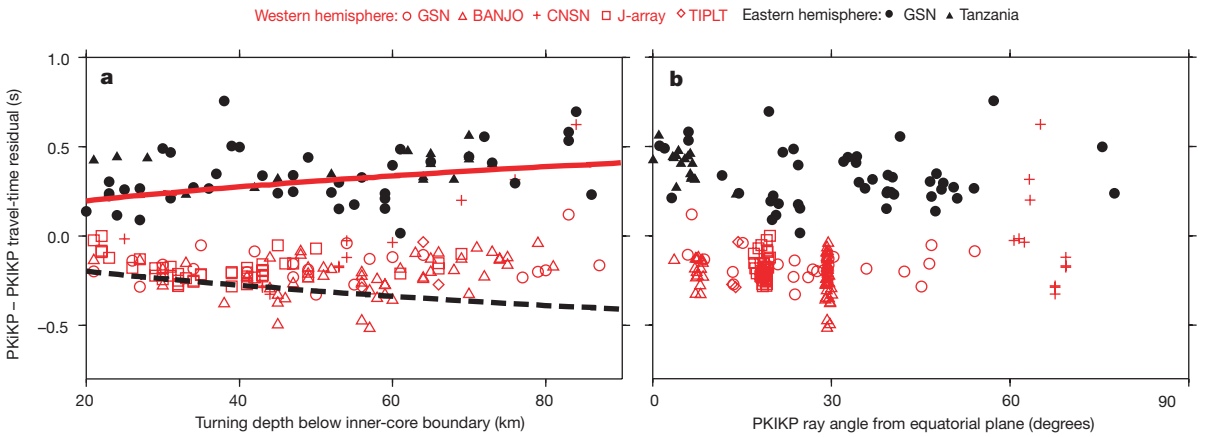


Figure 2 PKiKP – PKIKP travel-time residuals as a function of turning depth and angle. Residuals as function of (**a**) PKIKP turning depth below the ICB and (**b**) PKIKP ray angle from the equatorial plane. Solid and open symbols represent the eastern and western hemispheres, respectively. The predicted PKiKP – PKIKP travel-time residuals for two models with a top 100-km layer 0.5% faster (solid line) and 0.5% slower (dotted line) than the PREM in the inner core are shown.

© 2001 Macmillan Magazines Ltd

almost the same as the data-sampling rate and is much less than the observed time difference between the two hemispheres. The uncertainty in picking the maximal amplitudes is ±0.10 s. We have excluded the data in the distances that are less than 130°, because PKiKP and PKIKP waveforms interfere with each other in this distance range.

The observed hemispherical distribution of PKiKP – PKIKP travel-time residual cannot be explained by variation of the inner-core radius. An increase or decrease of the inner-core radius at the piercing points of PKIKP rays and the reflection points of PKiKP rays (Fig. 1 inset) would have almost no effect on PKiKP – PKIKP time. Small-scale topography, which may affect these two phases differently, is also an unlikely explanation for our observation, because the piercing points of PKIKP rays and the reflection

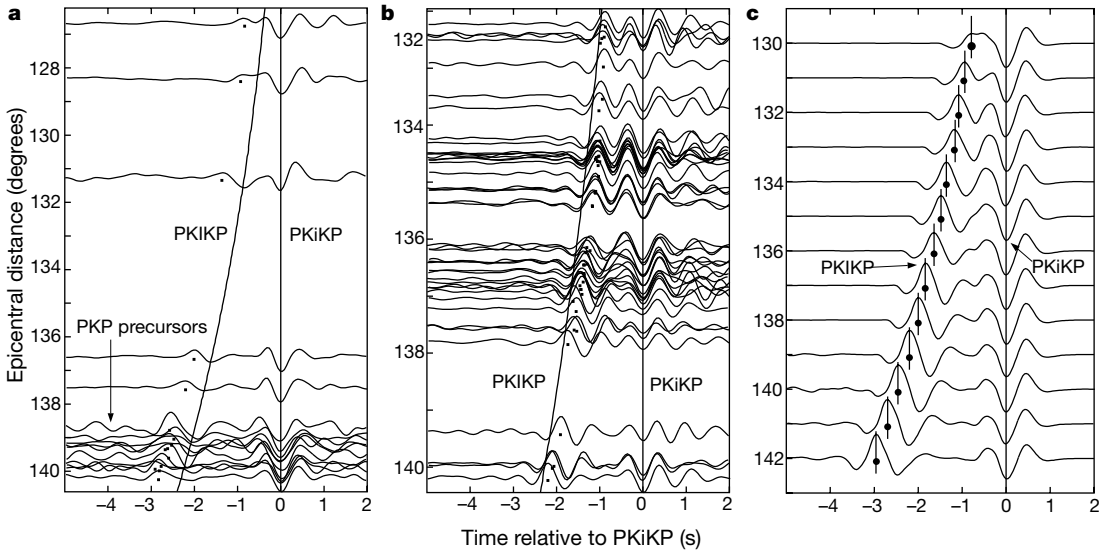


Figure 3 Examples of seismograms. Experimental seismograms were recorded in the Tanzania array (**a**) and the J-array (**b**), and synthetic seismograms were calculated by the generalized ray theory (**c**). All the seismograms are aligned according to the PKiKP arrivals. Observed and PREM[13] travel times of PKIKP are indicated by dots and straight line, respectively. **a**, 18 December 1994 at a source depth of 551.0 km and with $M_w = 5.7$. **b**, 26 April 1999 at a source depth of 174.0 km with $M_w = 5.9$. **c**, PKIKP resulted from waveform handpicking and ray tracing are indicated by vertical lines and dots, respectively.

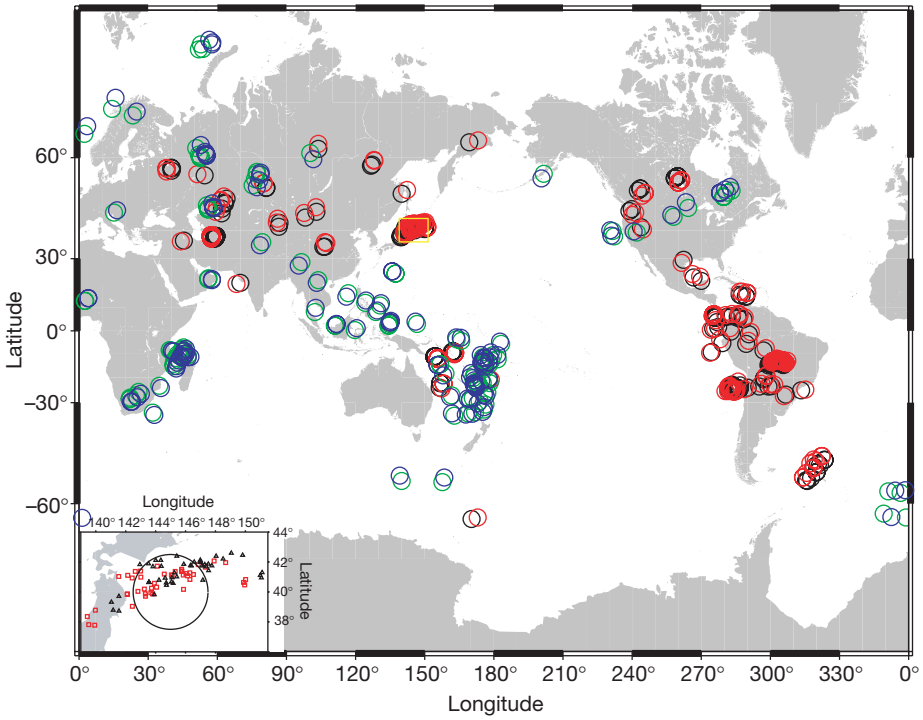


Figure 4 Distribution of hit points of PKIKP and PKiKP rays at the CMB. Black and red circles represent those of PKIKP and PKiKP bouncing at the western hemisphere of the top of the inner core; blue and green circles are those bounding at the eastern hemisphere. The size of the circles is roughly equal to the Fresnel zone of PKIKP and PKiKP at the CMB for a 1-Hz wave. The hit points within the yellow rectangle region, which are core exits of PKIKP and PKiKP observed by the J-array for the event 04/26/99, are shown in the inset. They are indicated by squares for PKIKP and triangles for PKiKP. A circle with a radius of 150 km is shown to indicate roughly the size of the Fresnel zone.

© 2001 Macmillan Magazines Ltd

points of PKiKP rays at the ICB overlap in some regions (such as the western Pacific Ocean).

The systematic variation of these differential travel times is also unlikely to be explained by the heterogeneities near the core–mantle boundary (CMB). In Fig. 4, we plot the hit points of PKIKP and PKiKP at the CMB using different colours, with black (PKIKP) and red (PKiKP) circles centring their bounding points in the western hemisphere and blue (PKIKP) and green (PKiKP) circles in the eastern hemisphere. The size of the circle is the same as the Fresnel zone of PKIKP and PKiKP at the CMB. The Fresnel zone is approximately 150 km at the CMB for a vertically propagating, short-period (about 1 Hz) P-wave. The separation between the PKiKP and PKIKP paths is, however, only about 50 km at the CMB. Because their Fresnel zones overlap (black and red circles; blue and green circles), the heterogeneities near the CMB would affect both phases in the same way. Thus the heterogeneities at the CMB would have little effect on the differential travel time of these two phases. Furthermore, some regions of the CMB (for example, Tonga, Europe; Fig. 4) are sampled by both rays bottoming in different hemispheres. In regions covered by dense seismic arrays, the PKiKP and PKIKP hit points overlap even for different observations. Detailed distribution of the hit points of PKIKP (red squares) and PKiKP (black triangles) observed by the dense J-array for the event on 26 April 1999 is shown in the inset of Fig. 4. The PKiKP – PKIKP travel-time residuals thus can only be attributed to the heterogeneity within the inner core.

In general, global PKiKP – PKIKP residual times are less scattered than PKP – PKIKP data, indicating that they are less biased by the seismic structures near the CMB and that our argument is reasonable. Our results are in good agreement with a PKP – PKIKP travel-time residual study[8], which first suggested a degree-one longitudinal distribution of heterogeneity in the top 100–500 km of the inner core. The PKP – PKIKP travel-time residuals are also affected by the structure of the uppermost 100 km of the inner core. Preliminary calculations suggest that our models can explain most of the PKP – PKIKP travel-time residuals.

The hemispherical distribution of heterogeneity in the top 100 km of the inner core may reflect some features of the inner-core growth. It may be caused either by an intrinsic difference during inner-core formation, or an external temperature difference at the base of the outer core. However, the former mechanism seems unlikely when we consider that the top 100 km of the inner core, as observed here, is clearly different from the rest of the inner core; the core structure is characterized by an axisymmetric anisotropic structure with a north–south fast axis[4–7]. Seismic velocity in the outermost inner core may be related to the temperature at the bottom of the liquid outer core, because the temperature will control the freezing rate of the liquid iron and therefore affects the seismic velocity of the solid iron. A recent experimental study[16] suggests that a cold downwelling at the CMB may generate a hemispherical variation of temperature above the ICB. The problems are whether and for how long the heterogeneity at the CMB can simply be represented by a higher-velocity region in the western Pacific. Therefore, the implication of the presence of a different layer in the outermost region of the inner core is very important to our understanding of the physical process of inner-core growth, and also of mantle dynamics. □

Received 18 July 2000; accepted 28 February 2001.

1. Yoshida, S., Sumita, I. & Kumazawa, M. Growth model of inner core coupled with the outer core dynamics and resulting elastic anisotropy. *J. Geophys. Res.* **101,** 28085–28103 (1996).
2. Anderson, O. L. Mineral physics of iron and of the core. *Rev. Geophys.* **33,** 429–441 (1995).
3. Boehler, R. Temperatures in the Earth's core from melting point measurements of iron at high static pressure. *Nature* **363,** 534–536 (1993).
4. Morelli, A., Dziewonski, A. M. & Woodhouse, J. H. Anisotropy of the core inferred from PKIKP travel times. *Geophys. Res. Lett.* **13,** 1545–1548 (1986).
5. Shearer, P. M., Toy, K. M. & Orcutt, J. A. Axi-symmetric Earth models and inner-core anisotropy. *Nature* **333,** 228–232 (1988).
6. Creager, K. C. Anisotropy of the inner core from differential travel times of the phases PKP and PKIKP. *Nature* **356,** 309–314 (1992).
7. Tromp, J. Support for anisotropy of the Earth's inner core from oscillations. *Nature* **366,** 678–681 (1993).
8. Tanaka, S. & Hamaguchi, H. Degree one heterogeneity and hemispherical variation of anisotropy in the inner core from PKP(BC) – PKP(DF) times. *J. Geophys. Res.* **102,** 2925–2938 (1997).
9. Creager, K. C. Large-scale variations in inner core anisotropy. *J. Geophys. Res.* **104,** 23127–23139 (1999).
10. Cormier, V. F. & Choy, G. L. A search for lateral heterogeneity in the inner core from differential travel times near PKP-D and PKP-C. *Geophys. Res. Lett.* **13,** 1553–1556 (1986).
11. Song, X. & Helmberger, D. V. Depth dependence of anisotropy of the Earth's inner core. *J. Geophys. Res.* **100,** 9805–9816 (1995).
12. Kaneshima, S. Mapping heterogeneity of the uppermost inner core using two pairs of core phases. *Geophys. Res. Lett.* **23,** 3075–3078 (1996).
13. Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Int.* **25,** 297–356 (1981).
14. Wen, L. & Niu, F. Seismic velocity and attenuation structures in the top of the Earth's inner core. *J. Geophys. Res.* (submitted).
15. Helmberger, D. V. in *Earthquakes: Observation, Theory and Interpretation* (ed. Kanomori, H.) 173–222 (Soc. Italiana Di Fisica, Bologna, 1983).
16. Sumita, I. & Olson, P. A laboratory model for convection in Earth's core driven by a thermally heterogeneous mantle. *Science* **286,** 1547–1549 (1999).
17. J-Array Group. The J-Array program: system and present status. *J. Geomag. Geoelectr.* **45,** 1265–1274 (1993).

Acknowledgements

We are grateful to the IRIS, J-array Data Center, W. Shannon and L. Saumure for supplying the data. We thank S. Tanaka for providing us their PKP-PKIKP data. We also thank H. Kawakatsu, S. Sacks and P. Silver for comments on the manuscript. Discussions with J. Aurnou, M. Fouch, W. Holt, S. Solomon, I. Sumita, S. Tanaka and D. Weidner were helpful in preparing the manuscript. This work is supported by the fellowship of the Carnegie Institution of Washington and an NSF grant.

Correspondence and requests for materials should be addressed to F.N. (e-mail: niu@dtm.ciw.edu).

The distribution of integumentary structures in a feathered dinosaur

Qiang Ji*, Mark A. Norell†, Ke-Qin Gao†, Shu-An Ji* & Dong Ren*

**Chinese Academy of Geological Sciences, 26 Baiwanzhuang, Beijing 100037, People's Republic of China*

† American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

Non-avian theropod dinosaurs with preserved integumentary coverings are becoming more common[1–6]; but apart from the multiple specimens of *Caudipteryx*, which have true feathers[2,7], animals that are reasonably complete and entirely articulated that show these structures in relation to the body have not been reported. Here we report on an enigmatic small theropod dinosaur that is covered with filamentous feather-like structures over its entire body.

The new specimen was collected from the extensive deposits of the Yixian Formation at Lingyuan, Liaoning, China. These fossil beds have produced many fossil specimens with preserved soft body structures[8–12]. Notable among these are fossil bird specimens[9,10]. Although some specimens from western Liaoning have been shown to be composites[13] or forgeries[14], the integrity of the specimen described here is assured because both slabs match up exactly and the integumentary covering lies below flakes of rock in several places: it is therefore not painted, scratched into the matrix or otherwise enhanced.

Theropoda Marsh 1881
Coelurosauria Huene 1914
Maniraptora Gauthier 1986
Dromaeosauridae Matthew & Brown 1922
Gen. et sp. indet.

© 2001 Macmillan Magazines Ltd

Locality. Fanzhangzi quarry, 20 km southwest of Lingyuan City, Liaoning Province, People's Republic of China.

Geological setting. The age of the fossil beds in western Liaoning is debated. Certain faunal elements suggest a Late Jurassic age[15]; radiometric work from several sites near Sihetun has suggested conflicting dates of 124.6 Myr ago[16] or 147 Myr ago[17]. The age of these beds is a complex problem and it is likely that several ages are represented at different quarry sites. No radiometric samples exist from the Fanzhangzi quarry and stratigraphic correlations are imprecise because the Fanzhangzi quarry is over 130 km from the Sihetun site.

Description. The specimen National Geological Museum of China (NGMC) 91 is preserved on a slab as part and counterpart (Fig. 1). Slab 91-A is the dorsal part, and slab 91-B the ventral part. The specimen is sprawled out with its head pointing towards the right side of the body. A small fish (*Lycoptera*) is preserved near the left foot. Judging from the size of the head in relation to the rest of the body it is probably a sub-adult[16,18], although the bones are very well ossified. The bones are brittle and most shattered when the specimen was collected by splitting the slab, so many skeletal details cannot be scored adequately. A few features allow us to place this animal confidently in the Maniraptora, the theropod group that includes birds and dromaeosaurid dinosaurs (notably the presence of a large semilunate carpal in the wrist; Fig. 2), and either within or closely related to the Dromaeosauridae, on the basis of the unique presence of elongate bifurcating prezygapophyses and chevrons that span several vertebrae[19].

NGMC 91 is similar to the dromaeosaurid *Sinornithosaurus*[4]. They have similar teeth. Both share several characters in common with dromaeosaurids, but they and *Microraptor*[5] display such uncharacteristic dromaeosaurid features as non-ginglymoid distal articular surfaces of the metatarsals and arctometatarsalian metatarsals. Some characters may prove diagnostic of this taxon (for example, the unusual bowed metacarpal I), but we cannot determine whether these are affected by ontogeny.

Because allometric scaling could account for some of the differences between *Sinornithosaurus* and NGMC 91, measurements from these skeletons were compared and found to be highly correlated ($r = 0.983$; see Supplementary Information). An allometric model for growth in the closely related *Archaeopteryx lithographica* has been developed[20]. This model shows that during growth the forelimb of *Archaeopteryx* disproportionately increases in length and the

Figure 1 NGMC 91-A, showing the distribution of filamentous integumentary structures across the entire body. Letters refer to details on Fig. 5.

Figure 2 The hands of NGMC 91-A, **a,** Left. **b**, Right, showing the presence of a distinct semilunate carpal.

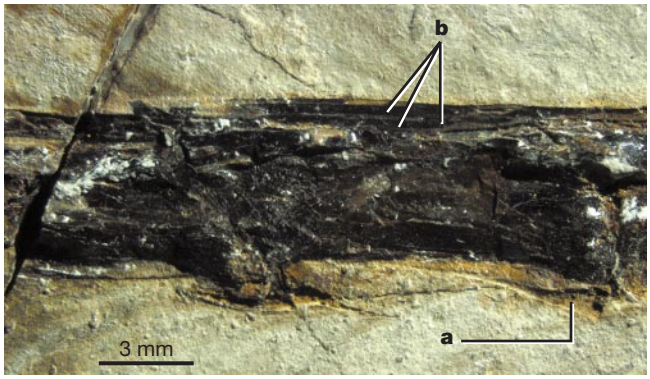


Figure 3 The tail of NGMC 91-A. **a**, Chevron with elongate anterior process. **b**, A bifurcation of prezygapophysis. Both features are diagnostic of Dromaeosauridae.

© 2001 Macmillan Magazines Ltd

skull decreases in length[20]. When measurements of NGMC 91 and *Sinornithosaurus* are incorporated into this model, a close fit is found. However, when compared with the 95% confidence interval for both of these taxa, the third phalanx of digit II and metacarpal II are significantly shorter, and metatarsal II is significantly longer than expected. Metatarsal II is the only tarsal element commonly preserved in all of the specimens, and only in NGMC 91 is the tibia preserved (where it is significantly longer than expected). If metatarsal II is reflective of total tarsal length, and the tibia of *Sinornithosaurus* is similar in proportion to NGMC 91, this indicates that the hindlimb of these animals is substantially longer than that of *Archaeopteryx*.

The skull is long and triangular in lateral view, more so than in other dromaeosaurids. The profile of the skull is concave between the orbit and the small elliptical nares. The orbit is large and filled with both superimposed rings of scleral ossicles. Posterior to the orbit lies a small, triangular postorbital. The lacrimal, which forms the preorbital bar, is T-shaped, as in other dromaeosaurids[21]. Inside the antorbital fossa, the vaulted anterior process of the palatine defining the choana can be observed, as is typical of dromaeosaurids[22] and troodontids[23]. No accessory antorbital fenestra is apparent.

Several dentary teeth can be observed. They are typically dromaeosaurid in that they are heavily recurved, not constricted at the base as in *Microraptor*[5], and strongly serrated with large blunt denticles posteriorly. There are no anterior serrations.

The cervical vertebrae are short and amphicoelous, and the thoracics are platycoelous. Apparently there were ten cervical vertebrae; the exact number of dorsals is difficult to determine. Uncinate processes and gastralia are present as reported for other dromaeosaurids[24,25]. Small rod-like bones in the chest area may represent ossified sternal ribs.

The anterior two-thirds of the tail is unusual because no individual vertebral segments are observable. Whether this is a preservational artefact or this section of the tail is composed of a single rod-like structure is difficult to determine. As in other dromaeosaurs[19], long stiffening rods surround the tail. They are composed of elongate, bifurcating prezygapophyses and chevrons (Fig. 3).

The pectoral girdle is poorly preserved. The thin hollow furcula is preserved as a boomerang-shaped element, similar to that of *Sinornithosaurus*[4] but much smaller. The left scapula is a long strap-like element with parallel sides.

The forelimbs are folded against the side of the torso. The humerus, radius and ulna are approximately the same length, and the hand is about one-third longer than these elements. The humerus is relatively narrow. Forelimb proportions relative to the femur are consistent with those previously reported for dromaeosaurids[2]. The ulna is bowed, as in other coelurosaurs[26], and approximately three times the diameter of the radius. There are no quill knobs on the ulna.

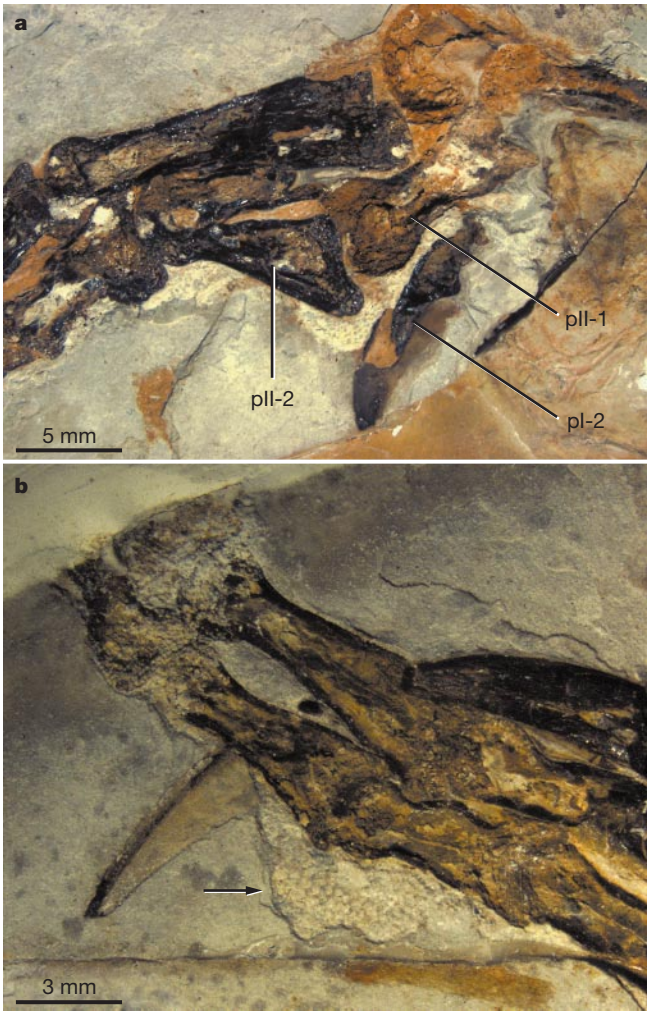


Figure 4 The pes of NGMC 91. **a**, The foot of NGMC 91-A showing the modified PII-1. **b**, Small patch of tuberculate impressions (arrow) on the foot, indicating skin impression.

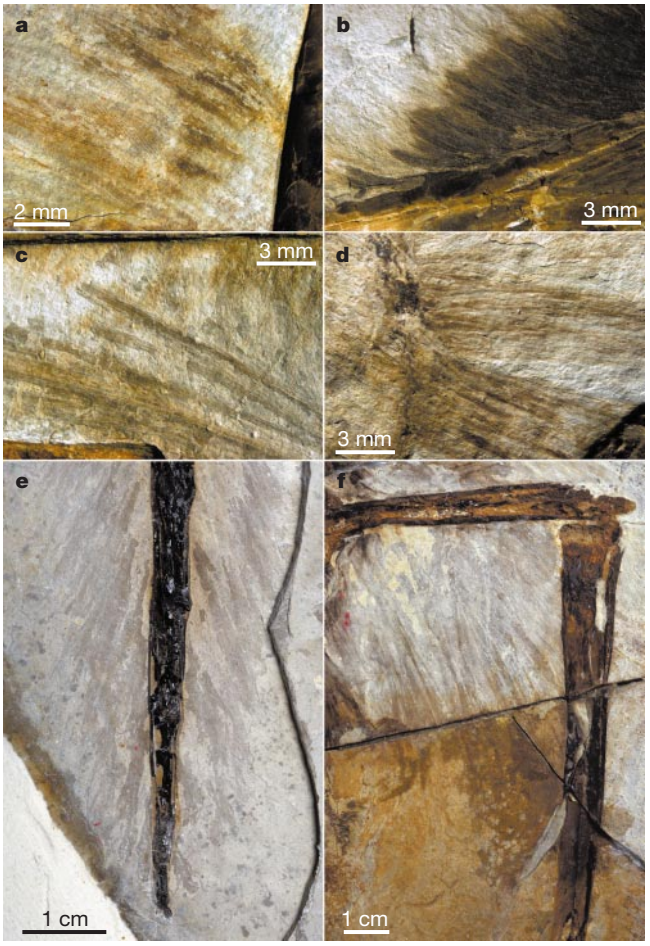


Figure 5 Integumentary structures of NGMC 91. **a**, Detail of branched integumentary structures on the posterior surface of the right ulna on NGMC 91-A. Note the tightly organized herring bone pattern. **b**, Detail of the integumentary filaments on the top of the head. **c**, **d**, Integumentary sprays on the shoulders of NGMC 91-A. **e**, Integumentary sprays surrounding the tail of NGMC 91-A. **f**, The long integumentary fibres on the left hindlimb.

© 2001 Macmillan Magazines Ltd

Although the wrist is poorly preserved, two proximal carpal elements are identifiable (Fig. 2). The semilunate carpal can be observed on the right wrist. Unusually, it appears to cap metacarpals II and III instead of I and II; however, this is likely to be a preservational artefact. An additional small, quadrangular carpal element just proximal to the semilunate, and lying at the terminus of the radius, represents the radiale.

The hand has the usual theropod configuration of three digits (digits I, II and III) and the typical phalangeal formula of 2-3-4. Metacarpals II and III are approximately equal in length. The unguals are not strongly recurved, but keratinous claw sheaths preserved on several of the digits indicate that the claws were recurved to over 180°. The unguals bear strong flexor tubercles. Digit II is the longest, as in nearly all theropod dinosaurs. Phalanges II-1 and II-2 are equal in length, and phalanx II-1 is extremely robust. The phalanges of digit III are very slender. Phalanx III-1 is short, being about half as long as phalanx II-1. As is typical of theropods, phalanx III-2 is very short and phalanx II-3 is about the same length as phalanx II-1.

The hindlimbs are long and gracile. The fibula contacts the ankle. The metatarsus is composed of three primary and two smaller elements. Metatarsal I is short and not retroverted. It lies extremely low on the side of metatarsal II, as in other dromaeosaurids[23,24]. Metatarsals II, III and IV are unfused. Metatarsals II and IV lie in apposition, indicating that MT-III may be pinched between them anteriorly, as in *Sinornithosaurus* and *Microraptor*[5], but different from other dromaeosaurids[19,24]. The phalangeal formula is ?2-3-4-5-0. The second pedal digit shows specializations of a raptorial digit and an enlarged claw, but not to the degree seen in dromaeosaurs such as *Velociraptor*. Surrounding the phalanges are small tuberculate impressions that probably represent skin impressions (Fig. 4). The unguals, like those of the hand, are short.

Sinornithosaurus[4], *Microraptor*[5] and NGMC 91 are preserved with patches of integumentary covering, although the full distribution of body covering on these taxa is unknown. In NGMC 91 integumentary filaments are distributed across the entire skull, body, limbs and tail, except for the distal hindlimbs (Fig. 5). Unlike *Sinosauropteryx*[6], the structure and distribution of filaments are heterogeneous. There are three basic types of filamentous structure: single fibres, long 'sprays' of fibres that resemble the plumulaceous feathers of *Protarchaeopteryx*[2], and fibres oriented around a central axis in a herring-bone pattern that resembles the remiges of *Caudipteryx*. Across the entire skeleton the filamentous covering is heavy and matted because filaments are preserved between several layers of sediment. In most cases the integumentary structures lie at around 90° to the skeletal elements.

Recent models of feather evolution based on development[27] predict that feathers went through evolutionary stages of unbranched ules, followed by the development of a rachis and finally barbs. The shape of the filaments on NGMC 91 meet these predictions, as several of these patterns are present on the specimen.

On the skull, loosely organized, oriented unbranched fibres are exposed along the midline (Fig. 4b). This is not a plume or crest, because several planes of matted filaments are separated by sediment. The entire head was apparently covered with a thick mat of filaments. These filaments extend a minimum of 2 cm from the skull. On the back of the neck, the filaments are longer, extending 3.2 cm from the neck and curving posteriorly.

Integument sprays organized around a central axis cover the shoulders and torso (Fig. 5c, d). These fibres are extremely thin and can be very long; on the hindlimb they are over 5 cm long (Fig. 5f). These sprays form a dense mat around the shoulders.

The most interesting integumentary areas are those on the arms and tail (Figs 5a, e and 6). In these areas the filaments are complex, bunched and tightly organized. Most of these filaments originate from a single point, forming a radiating spray. On the arms, distally a tight herring-bone pattern appears to be organized around a central rachis (Figs 5 and 6). This is the same as the pattern seen in *Caudipteryx zoui*[2]. Filaments emanate from the posterior edge of the ulna. These filaments are very long and uniformly perpendicular to the bone. The integumentary covering appears much flatter and more regular in this area; the filaments extend 5.2 cm from this element. The herring-bone appearance of the preserved impressions indicates that some organizing pattern, such as the barbules of modern birds, must have been present.

Oriented filaments are also distributed along the tail. Relatively short fibres lie at an angle of approximately 45° from the tail axis posteriorly. At the distal end of the tail the filaments become longer,

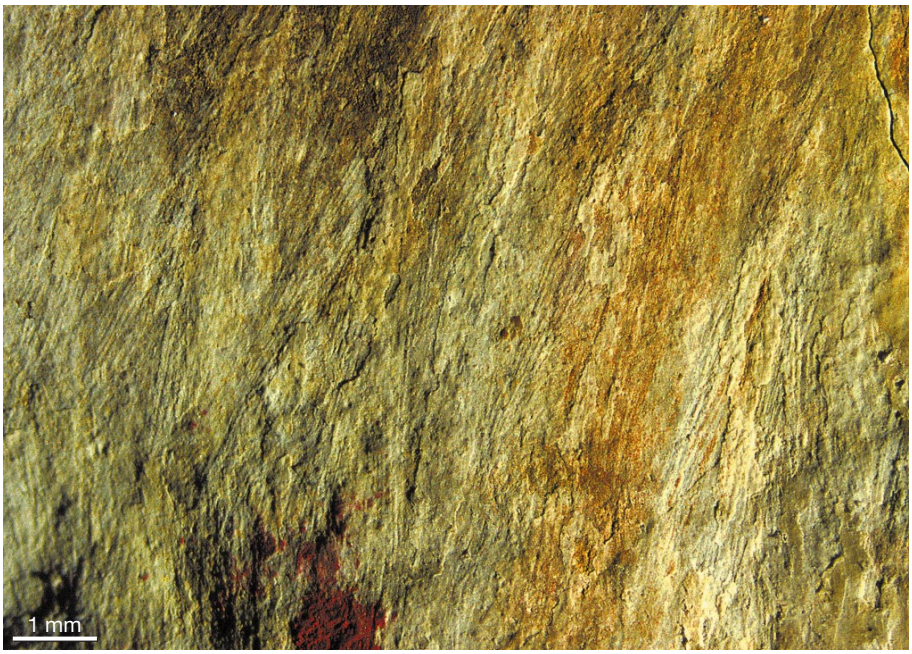


Figure 6 Detail of the herring bone pattern of tightly organized fibres lying adjacent to the posterior surface of the right ulna on NGMC 91-A.

© 2001 Macmillan Magazines Ltd

making the end lozenge-shaped.

Because feathers are the only integumental covering in vertebrates that have a tufted or branched structure[27] the occurrence of similar structures in NGMC 91, coupled with its phylogenetic position near the base of birds, is strong evidence that these structures are feather homologues. The myriad findings of flightless dinosaurs from Liaoning with similar integumentary structures that have been shown by independent phylogenetic studies[28,29] to be outside of the Avialae provide important evidence that the origin of feathers is unrelated to the origin of flight in Avialae. □

Received 29 November 2000; accepted 16 February 2001.

1. Ji, Q. & Ji, S. On discovery of the earliest bird fossil in China and the origin of birds. *Chin. Geol.* **17,** 30–33 (1996).
2. Ji, Q., Currie, P. J., Ji, S. & Norell, M. A. Two feathered dinosaurs from northeastern China. *Nature* **393,** 753–761 (1998).
3. Xu, X., Tang, Z. & Wang, X. A therizinosaurid dinosaur with integumentary structures from China. *Nature* **399,** 350–354 (1999).
4. Xu, X., Wang, X. & Wu, X. A dromaeosaurid dinosaur with a filamentous integument from the Yixian Formation of China. *Nature* **401,** 262–266 (1999).
5. Xu, X., Zhou, Z. & Wang, X. The smallest known non-avian theropod dinosaur. *Nature* **408,** 705–708 (2000).
6. Chen, P.-J., Dong, Z.-M. & Zhen, S.-N. An exceptionally well preserved theropod dinosaur from the Yixian Formation of China. *Nature* **391,** 147–152 (1998).
7. Zhou, Z.-H., Wang, X.-L., Zhang, F.-C. & Xu, X. Important features of *Caudipteryx*-evidence from two nearly complete new specimens. *Vert. PalAsiat.* **38,** 241–254 (2000).
8. Ji, Q., Luo, Z.-X. & Ji, S.-A. A Chinese triconodont mammal and mosaic evolution of the mammalian skeleton. *Nature* **398,** 326–330 (1999).
9. Hou, L. & Chen, P. *Liaoxiornis delicatus* gen. et sp. nov., the smallest Mesozoic bird. *Chin. Sci. Bull.* **44,** 311–314 (1999).
10. Hou, L. *Mesozoic Birds of China* (Phoenix Valley Provincial Aviary of Taiwan, 1997).
11. Gao, K., Tang, Z. & Wang, X. A long-necked diapsid reptile from the Upper Jurassic/Lower Cretaceous of Liaoning Province, northeastern China. *Vert. PalAsiat.* **37,** 1–8 (1999).
12. Gao, K., Evans, S., Ji, Q., Norell, M. & Ji, S. Exceptional fossil material of a semi-aquatic reptile from China: the resolution of an enigma. *J. Vert. Paleontol.* **20,** 417–421 (2000).
13. Simmons, L. M. Archaeoraptor fossil trail. *Nat. Geogr. Mag.* October 128–132 (2000).
14. Chiappe, L. M., Ji, S. A., Ji, Q. & Norell, M. A. Anatomy and systematics of the Confuciusornithidae (Theropoda: Aves) from the Late Mesozoic of northeastern China. *Bull. Am. Mus. Nat. Hist.* **242,** 1–89 (1999).
15. Ji, Q. *et al.* On the sequence and age of the protobird-bearing deposits in the Sihetun-Jianshangou area, Beipiao, western Liaoning. *Prof. Papers Stratigr. Paleontol.* **27,** 74–80 (1999).
16. Swisher, C. C., Wang, Y.-Q., Wang, X.-L., Xu, X. & Wang, Y. Cretaceous age for the feathered dinosaurs of Liaoning, China. *Nature* **400,** 58–61 (1999).
17. Lo, C.-H., Chen, P.-J., Tsou, T.-Y., Sun, S.-S. & Lee, C.-Y. in *Jehol Biota* (eds Chen, P.-J. & Jin, F.) 328–340 (Science Press, Beijing, 1999).
18. Varricchio, D. J. in *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 749–754 (Academic, San Diego, 1997).
19. Ostrom, J. H. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 269–279 (Univ. California Press, Berkeley, 1990).
20. Houck, M. A. *et al.* Allometric scaling in the earliest fossil bird. *Archaeopteryx lithographica. Science* **247,** 195–198 (1990).
21. Currie, P. J. New information on the anatomy and relationships of *Dromaeosaurus albertaensis* (Dinosauria: Theropoda). *J. Vert. Paleontol.* **15,** 576–591 (1995).
22. Barsbold, R. & Osmólska, H. The skull of *Velociraptor* (Theropoda) from the Late Cretaceous of Mongolia. *Acta Palaeontol. Polonica* **44,**189–219 (1999).
23. Norell, M. A., Makovicky, P. J. & Clark, J. M. A new troodontid theropod from Ukhaa Tolgod, Mongolia. *J. Vert. Paleontol.* **20,** 7–11 (2000).
24. Norell, M. A. & Makovicky, P. J. Important features of the dromaeosaurid skeleton: information from a new specimen. *Am. Mus. Novit.* **3215,** 1–28 (1997).
25. Norell, M. A. & Makovicky, P. J. Important features of the dromaeosaurid skeleton II: information from newly collected specimens of *Velociraptor mongoliensis*. *Am. Mus. Novit.* **3282,** 1–45 (1999).
26. Gauthier, J. in *The Origin of Birds and the Evolution of Flight* (ed. Padian, K.) 1–55 (California Academy of Sciences, San Francisco, 1986).
27. Prum, R. O. The development and evolutionary origin of feathers. *J. Exp. Zool.* **285,** 291–306 (1999).
28. Sereno, P. C. The evolution of dinosaurs. *Science* **284,** 2137–2147 (1999).
29. Holtz, T. R., Jr A new phylogeny of the carnivorous dinosaurs. *Gaia* **15,** 5–61 (2000).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

Comments from J. Clark, P. Makovicky, K. Padian, R. Prum and M. Siddall improved the manuscript. M. Ellison provided the Figures. We thank X. Xing and Z. Zhonghe for access to *Sinornithosaurus*. Support was provided by the Division of Paleontology (AMNH); the Ministry of Land Resources of the People's Republic of China; the National Natural Science Foundation of China; the National Geographic Society; R. Byron and L. Jaffe; and V. Pan.

Correspondence and requests for materials should be addressed to M.A.N. (e-mail: norell@amnh.org).

Genetic evidence for Near-Eastern origins of European cattle

Christopher S. Troy*†, David E. MacHugh‡†, Jillian F. Bailey*, David A. Magee*, Ronan T. Loftus*, Patrick Cunningham*, Andrew T. Chamberlain§, Bryan C. Sykes‖ & Daniel G. Bradley*

* *Department of Genetics, Smurfit Institute, Trinity College, Dublin 2, Ireland*
‡ *Department of Animal Science and Production, Faculty of Agriculture, University College Dublin, Belfield, Dublin 4, Ireland*
§ *Department of Archaeology and Prehistory, University of Sheffield, Sheffield S1 4ET, UK*
‖ *Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK*
† *These authors contributed equally to this work*

The limited ranges of the wild progenitors of many of the primary European domestic species point to their origins further east in Anatolia or the fertile crescent[1,2]. The wild ox *(Bos primigenius)*, however, ranged widely[3] and it is unknown whether it was domesticated within Europe as one feature of a local contribution to the farming economy[1,2,4]. Here we examine mitochondrial DNA control-region sequence variation from 392 extant animals sampled from Europe, Africa and the Near East, and compare this with data from four extinct British wild oxen. The ancient sequences cluster tightly in a phylogenetic analysis and are clearly distinct from modern cattle. Network analysis of modern *Bos taurus* identifies four star-like clusters of haplotypes, with intra-cluster diversities that approximate to that expected from the time depth of domestic history. Notably, one of these clusters predominates in Europe and is one of three encountered at substantial frequency in the Near East. In contrast, African diversity is almost exclusively composed of a separate haplogroup, which is encountered only rarely elsewhere. These data provide strong support for a derived Near-Eastern origin for European cattle.

A deep bifurcation in bovine mitochondrial DNA (mtDNA) phylogeny has been described[5,6] and is indicative of a pre-domestic divergence well in excess of 100,000 years between the two cattle taxa, *Bos indicus* and *B. taurus.* Both clades were observable in these data; 383 samples were of *B. taurus* mtDNA type whereas 9 may be classed as variants of *B. indicus.* The latter were encountered as minority types in five morphologically taurine Near-Eastern populations, and fall within Indian cattle haplotypes in the left-hand cluster of the phylogeny (Fig. 1). The aurochs sequences are more closely related to modern *B. taurus*, but it is significant that they cluster tightly in isolation, well outside the range of sequence variation among extant taurine. This phylogenetic consistency is an indication of authenticity of these ancient data. This is shared with two sequences[7] (analysed in a separate laboratory), which also display the eight transitions that separate each wild ox sequence from the modern *B. taurus* root sequences. Additionally, each sequence has been confirmed in repeated extractions from several samples and was ascertained directly from polymerase chain reaction (PCR) products, without a cloning step.

Within the 383 *B. taurus* mtDNA sequences examined, 152 haplotypes were identified that were defined by 77 polymorphic sites. Among the 152 haplotypes detected, the most predominant (T3) occurs 99 times; T1 occurs 39 times; T2 occurs eight times; another two occur seven times; two occur six times; three (including haplotype T) occur five times; two occur four times; 14 occur in triplicate; 20 are duplicated; and 106 haplotypes are unique.

Genetic loci from a centre of origin are expected to retain more ancestral variation and show higher haplotypic and nucleotide diversity, with lineage pruning through successive colonization events leading to a reduction in derived populations. Accordingly,

© 2001 Macmillan Magazines Ltd

diversity is visibly highest here within breeds from the Middle East and Anatolia (Fig. 2). Mean pairwise differences observed between 240 base-pair (bp) haplotypes from the Middle East are 3.97 (s.d. = 2.03) and 3.49 (s.d. = 1.81) from Anatolia. Values for the European mainland (1.92; s.d. = 1.10), Britain (2.68; s.d. = 1.45), western-fringe Europe (1.47; s.d. = 0.91) and Africa (2.09; s.d. = 1.18) are consistently lower.

On examination of reduced median networks of the 383 *B. taurus* mtDNA sequences (Fig. 2) we found that almost all sequences root back to the phylogeny through one of four main haplotypes that are also numerically prominent. Figure 2 includes a skeleton network that estimates their phylogenetic relationship. These main haplotypes are designated as T1, T2 and T3, which coalesce to the central sequence T—identified as the most probable *B. taurus* root sequence when the central *Bos primigenius* haplotype is added to the skeleton network (not shown).

The numerical and topological predominance of these four haplotypes suggests that they are ancestral; furthermore, the star-like pattern of derived variants surrounding each haplotype is consistent with a history of population expansion[8]. Two other treatments of the data support this demographic model. First, each geographical grouping yields smooth pairwise mismatch distributions that are suggestive of expansions[9]. Second, Fu's *F*s statistic, which is particularly sensitive to population growth[10,11], yields a highly significant departure from neutrality in each sample ($P < 0.0001$).

In Fig. 2 all sequences are colour coded into four haplogroups, indicating which root haplotype they coalesce to. Notably, these haplogroups are geographically distributed. Both the Anatolian and Middle-Eastern networks are primarily composed of three, star-like phylogenies that are centred on haplotypes T, T2 and T3. The expansion signatures of these haplotypes within the putative centre of domestication may indicate that multiple primary haplotypes were established during this phase of aurochs capture. Members of the fourth haplogroup, T1, occur at a relatively low frequency in Anatolian and Middle-Eastern regions (twice and three times, respectively) but predominate almost absolutely in Africa where 89 out of 95 sequences coalesce together in a markedly star-like phylogeny. This haplogroup is absent from the European regions presented here, although African sequences have been reported in cattle from southern Portugal[12] where they are probably a legacy of North African historical influence. African haplotypes have also been documented in populations from the Americas[13]—perhaps as a secondary consequence of slave trade routes.

In a marked display of symmetry with the African pattern, each regional European sample produces an essentially star-like phylogeny centred, in this case, on sequence T3. Such phylogenetic topologies are suggestive of past population expansions and estimates of time depths to such events may be made using ρ, the average mutational distance from the central sequence in a haplogroup, and an estimate of mutation rate (38% per Myr)[14]. Of note, the 95% CI range of estimates for expansions around the four haplogroup clusters are consistent with the time depth of cattle domestication (T, T1, T2 and T3 yield 5,600–21,100, 4,400–16,500, 5,500–20,700 and 4,300–16,100 yr BP, respectively). Under the assumption that the advent of agropastoralism would have resulted in dramatic and sustained population increases of both the early herders and their flocks, these patterns are probably genetic signatures of population expansion after domestication.

The relationship between African haplotypic variation and that of the Near East seems to be qualitatively different from that between the latter and Europe. Most of the African diversity is clustered around a haplotype that is absent from our European samples and encountered only at very low frequency in Anatolia and the Near East. This observation and the pattern and extent of this T1 cluster diversity suggest that it is the result of a domestication-induced expansion. The near absence of this haplogroup in the Middle East and Anatolia suggests that the expansion did not occur within the fertile crescent. This provides some support for archaeological assertions of a separate African domestication[15]. In contrast, reduced diversity and the predominance of one haplogroup T3, which is also encountered as a subset of the variation encountered in Anatolia and the Middle East, strongly suggests that European cattle are derivatives of the Near-Eastern Neolithic complex. This evidence, together with the marked phylogenetic distinction of British aurochs sampled over a wide time depth, does not support the hypothesis that local domestication made any significant contribution to the establishment of agropastoralism in Europe.

Notably, whereas African variants coalesce to a different root haplotype, the difference between that and the other *B. taurus* ancestral sequences is small. The overall coalescence of all sampled domestic *B. taurus* mtDNA sequences around the central sequence T is estimated as 10,100–37,600 yr BP. This shallow divergence may indicate an ultimate origin for all extant *B. taurus* in an ancestral

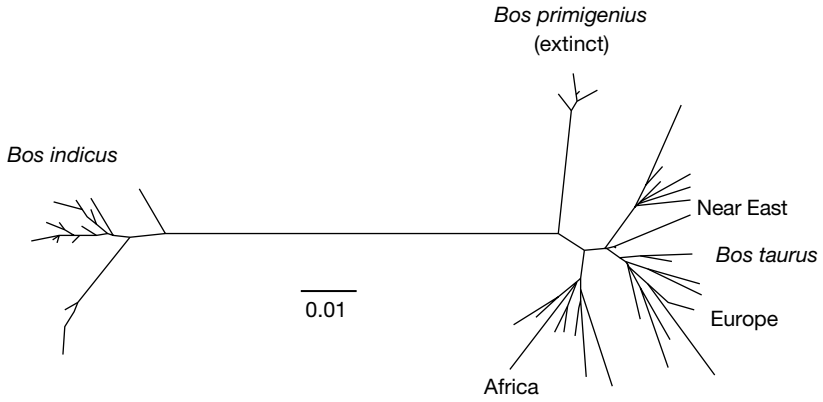


Figure 1 Neighbour-joining phylogeny of *B. indicus, B. taurus* and extinct British *B. primigenius* mtDNA control-region haplotypes. The sequences compared are 201 bp in length and include positions 16,042–16,158 and 16,179–16,262. Gaps were excluded and no distance correction was employed. The four ancient sequences cluster tightly (bootstrap value of 94%; 1,000 replications) and form an outgroup to the modern *B. taurus* haplotypes. This is despite their origin in animals dispersed geographically and chronologically. These ancient sequences all share eight transitions that separate them from the modern *B. taurus* root sequence. Notably, these substitutions are also shared with two previous British aurochs sequences[7]. The *B. indicus* cluster yielded a bootstrap value of 100% and the phylogeny root, estimated using bison as an outgroup (data not shown), occurs between this clade and all others. A scale bar (divergence of 0.01) is shown.

© 2001 Macmillan Magazines Ltd

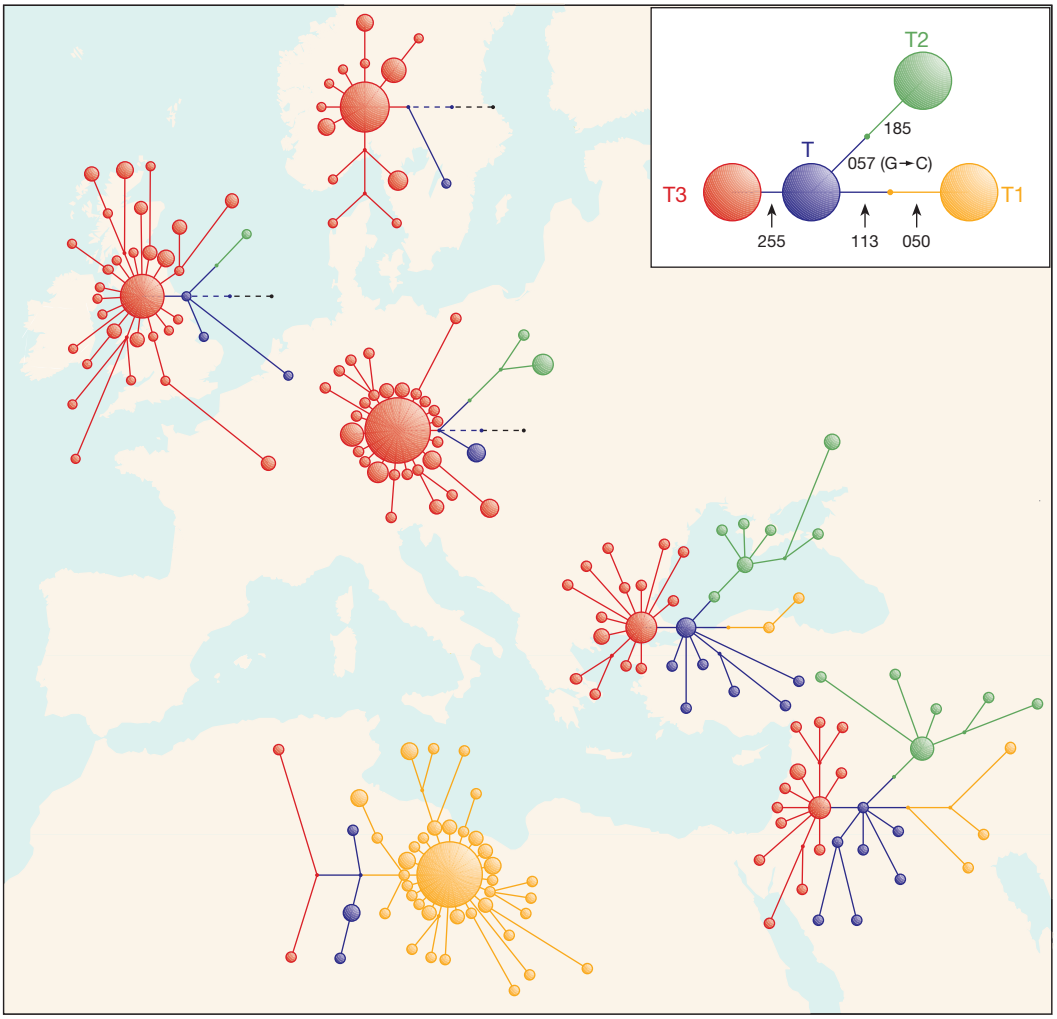


Figure 2 *B. taurus* mtDNA reduced median networks constructed from six regional haplotype groups. Inset, The relationships of the four central, primary *B. taurus* haplotypes, T, T1, T2 and T3. T is defined by a transition at position 16,255 from the reference sequence[18]; T1 by transitions at 16,050, 16,113 and 16,255; T2 by transitions at 16,185 and 16,255 plus a G to C transversion at 16,057; and T3 is identical to the reference sequence. The spatial arrangement of the skeleton network and the colour codes are preserved in the full data networks (placed in the region of origin on the background map). The samples are grouped as originating in the Middle East, Anatolia, mainland Europe, Britain, western-fringe Europe and Africa. Reduced median phylogenetic networks were constructed manually. Haplotypes encountered in each region (coloured circles) and unsampled intermediate nodes or unsampled primary haplotypes (small points) are shown. Circle areas are proportional to haplotype frequencies and colour coding indicates which of the four skeleton network haplotypes they root to. Most (71%) sites that were reduced correspond to the hypermutable sites identified above, and are underlined below. Networks were reduced at the following positions: Anatolian sequences, 16,050, 16,057, 16,074, 16,110, 16,113, 16,138, 16,247 and 16,248; Middle-Eastern sequences, 16,049, 16,050, 16,058, 16,074, 16,085, 16,109, 16,113, 16,121, 16,122, 16,200, 16,231, 16,247 and 16,260; Continental Europe sequences, 16,110, 16,164 and 16,260; British sequences, 16,049, 16,050, 16,057, 16,074, 16,109, 16,113, 16,122, 16,127 and 16,138. The Western Europe network did not require resolving. Geographical distribution of the four haplogroups is clear. T3 predominates in Europe and along with T and T2 comprises almost all Near-Eastern variation. Haplogroup T1 dominates African diversity but is scarcely represented elsewhere.

population of the wild ox, which was itself limited in diversity, perhaps as a result of climatically mediated range restriction. □

Methods

Modern samples

We examined 392 animals from 34 breeds: 208 from Europe, 89 from the Near East and 95 from Africa. For network analysis, populations were subdivided into six geographical groupings: western-fringe Europe (Icelandic, Kerry, Norwegian Red, Telemark, Westland Fjord); Britain (Galloway, Highland, Aberdeen Angus, Hereford, Jersey); European mainland (German Black, Friesian, Limousin, Charolais, Simmental, Romagnola, Barrenda, Sykia); Anatolia (Anatolian Black, East Anatolian Red, South Anatolian Red, Turkish Grey); Middle East (Damascus, Kurdish, North Iraq, South Iraq); and Africa (Egyptian, N'Dama, Somba, Kapsiki, Namchi, Kuri, Butana, Kenana, White Fulani).

Archaeological samples

The six aurochs mtDNA sequences examined as part of this survey consisted of two previously published[7] sequences, D740 and D812, and four new sequences from *B. primigenius* skeletal remains excavated from chronologically distinct sites widely dispersed across present-day England. The four new samples are (sample code and estimated age in brackets): Charterhouse Warren Farm Swallet (CHWF, 4,090–3,720 BP); Totty Pot (TP65, 7,570–7,320 BP); Carsington Pasture Cave (CPC98, 6,200–5,650 BP); and North Ferriby (NORF, 3,990–3,720 BP).

Extraction, DNA amplification and sequencing

DNA samples were extracted from blood[5], semen[13] and hair (see below). We used about 10–12 hair follicles per animal. Follicles were treated with 50 μl of 200 mM NaOH at 97 °C for 30 min. Then 50 μl of 200 mM HCl, 100 mM Tris-HCl, pH 8.5 was added. PCR and sequencing conditions were as described[12]. The region analysed was a defined, highly variable region of the mtDNA control region between bases 16,023 and 16,262 (ref. 5). We extracted and purified archaeological samples essentially as described[16]. General ancient DNA laboratory practices, PCR amplification procedures and DNA sequencing methods have been detailed before[17]. Two overlapping PCR products were amplified from each extract ($AN2_{FOR}$–$AN1_{REV}$ and $AN1_{FOR}$–$AN3_{REV}$). The PCR primers and their location and orientation in the reference sequence[18] are: $AN2_{FOR}$ (16,022–16,041); $AN1_{REV}$ (16,178–16,159); $AN1_{FOR}$ (16,159–16,178); and $AN3_{REV}$ (16,334–16,314). At least two independent samples were taken from each bone in the panel. DNA was extracted and purified from each independent sample in duplicate. A range of PCR amplifications and

© 2001 Macmillan Magazines Ltd

subsequent direct sequence assays were performed from each purified DNA extract. The mtDNA sequences derived from each bone were therefore multiply verified through independent samples, extractions, amplifications and sequence determinations. In all cases for each bone in the panel, the replicated mtDNA sequences were consistent across all samples and extractions. The veracity and integrity of the aurochs sequence haplogroup is strongly supported by the congruity of the bones analysed in Dublin (CHWF, TP65, CPC98 and NORF) with those previously examined in Oxford (D740 and D812).

Analysis of mitochondrial sequence data

We aligned sequences by eye. Data from previous studies[5–7,17] were also included and novel sequences were submitted to GenBank (accession numbers AF336383–AF3366748). We constructed a neighbour-joining phylogeny using uncorrected distances with gapped positions excluded and the Neighbour program in the Phylip package[19]. Arlequin 2.0 (ref. 20) was used to estimate nucleotide diversity values for each population, to estimate parameters and goodness of fit from mismatch distributions and to compute Fu's *Fs* test of selective neutrality[10]. Phylogenetic analysis was also performed using reduced median networks, constructed manually according to ref. 21. Sites prone to hypermutablity were identified using a network of zebu and taurine samples examined in ref. 6. For the reduced median networks shown, most of the reductions (71%) were at these sites. The mutation rate for the 240-bp region of the cattle D-loop was calculated using the estimate of the transition/transversion ratio for the sequence data presented here (61/1). This agrees well with previous estimates on smaller data sets 57/1 (ref. 6) and 41/1 (ref. 7), and as the new rate incorporates data presented in those papers it is used in preference to either. The *Bison–Bos* divergence is considered to be around 1 Myr. The three transversions observed in the 240-bp region between the *Bison–Bos* groups constitute the equivalent of 183 transitions. The one-lineage rate was estimated as 38% per Myr or one substitution per 10,928 yr. The central 95% credible region for the expansion time was calculated for each of the four main cattle haplogroups using the program CRED[22].

Received 10 August 2000; accepted 19 January 2001.

1. Smith, B. D. *The Emergence of Agriculture* (Scientific American Library, New York, 1995).
2. Harris, D. R. in *The Origins and Spread of Agriculture and Pastoralism in Eurasia* (ed. Harris, D. R.) (UCL, London, 1996).
3. Epstein, H. *The Origin of the Domestic Animals of Africa* (Africana, New York, 1971).
4. Bökönyi, S. *The History of Domestic Animals in Central and Eastern Europe* (Akadémiai Kiadó, Budapest, 1974).
5. Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M. & Cunningham, P. Evidence for two independent domestications of cattle. *Proc. Natl Acad. Sci. USA* **91,** 2757–2761 (1994).
6. Bradley, D. G., MacHugh, D. E., Cunningham, P. & Loftus, R. T. Mitochondrial diversity and the origins of African and European cattle. *Proc. Natl Acad. Sci. USA* **93,** 5131–5135 (1996).
7. Bailey, J. F. *et al*. Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor species. *Proc. R. Soc. Lond. B* **263,** 1467–1473 (1996).
8. Richards, M. B., Macaulay, V. A., Bandelt, H.-J. & Sykes, B. C. Phylogeography of mitochondrial DNA in western Europe. *Ann. Hum. Genet.* **62,** 241–260 (1998).
9. Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9,** 552–569 (1992).
10. Fu, Y.-X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147,** 915–925 (1997).
11. Excoffier, L. & Schneider, S. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proc. Natl Acad. Sci. USA* **96,** 10597–10602 (1999).
12. Cymbron, T., Loftus, R. T., Malheiro, M. I. & Bradley, D. G. Mitochondrial sequence variation suggests an African influence in Portuguese cattle. *Proc. R. Soc. Lond. B* **266,** 597–603 (1999).
13. Troy, C. S. *Mitochondrial DNA Phylogeny and Biogeography of Domestic Cattle.* Ph.D Thesis, Univ.-Dublin (1998).
14. Macaulay, V. *et al*. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.* **64,** 232–249 (1999).
15. Wendorf, F. & Schild, R. Nabta Playa and its role in northeastern African prehistory. *J. Anthropol. Archaeol.* **17,** 97–123 (1998).
16. Yang, D. Y., Eng, B., Waye, J. S., Dudar, J. C. & Saunders, S. R. Technical note: improved DNA extraction from ancient bones using silica-based spin columns. *Am. J. Phys. Anthropol.* **105,** 539–543 (1998).
17. MacHugh, D. E. *et al*. Early mediaeval cattle remains from a Scandinavian settlement in Dublin: genetic analysis and comparison with extant breeds. *Phil. Trans. R. Soc. Lond. B* **354,** 99–109 (1999).
18. Anderson, S. *et al*. Complete sequence of bovine mitochondrial DNA. *J. Mol. Biol.* **156,** 683–717 (1982).
19. Felsenstein, J. PHYLIP: Phylogeny Inference Package. (University of Washington, Seattle, 1993).
20. Schneider, S., Roessli, D. & Excoffier, L. *ARLEQUIN Ver. 2.0: a Software for Population Genetics Data Analysis* (Department of Anthropology, University of Geneva, Switzerland, 1999).
21. Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. Mitochondrial portraits of human populations using median networks. *Genetics* **141,** 743–753 (1995).
22. Macaulay, V. A. CRED: Credible Regions for Coalescence Times. (Univ. Oxford, 1998).

Acknowledgements

We thank C. Hawkes, O. Ertugrul, A. H. Al Haboby, A. H. Harba, M. A. A. El-Barody, E. Thompson, T. Goodchild, H. Halila, A. Swaid, G. Guneren, B. Tekbas, M. Bruford, B. Sauveroche, G. Kana, D. Achu-Kwi, M. Diallo, L. Gnaho, K. Papadopolous, A. G. Georgoudis, C. Gaillard, O. Hanotte, E. Rege, the Nordic GeneBank and C. Hawes for assistance or provision of samples. We also thank the Highland Cattle Society and the Black and Belted Galloway Societies for sample information. This work was partly funded by a European Commission contract. D.E.M. was supported by a Wellcome Trust Fellowship in Bioarchaeology. J.F.B. is a European Commission Marie Curie Fellow. Radiocarbon dating at the Oxford Radiocarbon Accelerator Unit was funded by NERC.

Correspondence and requests for materials should be addressed to D.G.B. (e-mail: dbradley@tcd.ie).

The highly reduced genome of an enslaved algal nucleus

Susan Douglas*, Stefan Zauner†, Martin Fraunholz†‡, Margaret Beaton‡§, Susanne Penny*, Lang-Tuo Deng§, Xiaonan Wu§, Michael Reith*, Thomas Cavalier-Smith‡§ & Uwe-G Maier†

* *National Research Council of Canada Institute for Marine Biosciences and Program in Evolutionary Biology, Canadian Institute of Advanced Research, 1411 Oxford Street, Halifax, Nova Scotia B3H 3ZI, Canada*
† *Cell Biology and Applied Botany, Philipps-University Marburg, Karl-von-Frisch-Strasse, D-35032 Marburg, Germany*
§ *Program in Evolutionary Biology, Canadian Institute of Advanced Research, Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada*

Chromophyte algae differ fundamentally from plants in possessing chloroplasts that contain chlorophyll *c* and that have a more complex bounding-membrane topology[1]. Although chromophytes are known to be evolutionary chimaeras of a red alga and a non-photosynthetic host[1], which gave rise to their exceptional membrane complexity, their cell biology is poorly understood. Cryptomonads are the only chromophytes that still retain the enslaved red algal nucleus as a minute nucleomorph[2–4]. Here we report complete sequences for all three nucleomorph chromosomes from the cryptomonad *Guillardia theta*. This tiny 551-kilobase eukaryotic genome is the most gene-dense known, with only 17 diminutive spliceosomal introns and 44 overlapping genes. Marked evolutionary compaction hundreds of millions of years ago[1,4,5] eliminated nearly all the nucleomorph genes for metabolic functions, but left 30 for chloroplast-located proteins. To allow expression of these proteins, nucleomorphs retain hundreds of genetic-housekeeping genes[5]. Nucleomorph DNA replication and periplastid protein synthesis require the import of many nuclear gene products across endoplasmic reticulum and periplastid membranes. The chromosomes have centromeres, but possibly only one loop domain, offering a means for studying eukaryotic chromosome replication, segregation and evolution.

Soon after the symbiogenetic origin of chloroplasts from cyanobacteria[1] to form the common ancestor of green plants, red and glaucophyte algae (kingdom Plantae[6,7]), even more complex eukaryotic cells arose by secondary symbiogenesis[1,3,4] (Fig. 1). Such chimaeric integration of two evolutionarily distant eukaryotic cells occurred independently in the common ancestor of cryptomonads and other chromophytes, in which the endosymbiont was a red alga, and in chlorarachneans, which acquired a green alga[1,3,4]. In both cryptomonads and chlorarachneans, a flagellate host contributed the nucleus, endomembranes and mitochondria to the chimaera, whereas the photosynthetic endosymbiont provided its chloroplast, plasma membrane (the periplastid membrane[1,3,4]) and a second nucleus (the nucleomorph), which became miniaturized[3–5]. The nucleomorph of both groups kept its envelope, nuclear pores[8] and three minute chromosomes[9]. In the ancestor of cryptomonads and chromobiotes (treated as kingdom Chromista[6,8]) but not chlorarachneans, the former food vacuole membrane originally enclosing the enslaved endosymbiont apparently fused with the nuclear envelope[10], placing it in the rough endoplasmic reticulum[8,10] (RER; Fig. 1). Cryptomonad cells depend on four genomes, each encoding distinct protein synthesis machineries in discrete

‡ Present addresses: Goddard Laboratories, University of Philadelphia, 415 South University Avenue, Philadelphia, Pennsylvannia 19104-6018, USA (M.F.); Biology Department, 63B York Street, Mount Allison University, Sackville, NB, Canada E4L 1G7 (M.B.); Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK (T.C.-S.).

© 2001 Macmillan Magazines Ltd

compartments, between which proteins are translocated. Until now, understanding how these genomes cooperate has been limited by the availability of only partial sequences for nucleomorph genomes[11,12].

We report the first complete genome sequence of a nucleomorph (551,264 base pairs (bp)), which proves conclusively that it is a vestigial nucleus[13] and that cryptomonads comprise one eukaryotic cell nested in another. Gene density is extremely high (1 gene per 977 bp) and non-coding regions are ultrashort, with only one pseudogene φ*rpl24* (Fig. 2). The six chromosome ends are identical repeats, comprising telomeres ($([AG]_7AAG_6A)_{11}$, 5S and 28S/5.8S/18S ribosomal RNA genes and five open reading frames (ORFs); the ubiquitin conjugation enzyme gene is repeated at five ends, and the TATA-box binding protein gene (*tfIId*) at three ends. Except in these repeats and five central 200–500-bp regions (possibly centromeric), intergenic spacers have few, if any, nucleotides. Forty-four genes overlap by up to seventy-six nucleotides (Fig. 2). One gene has three copies, one on chromosome 1 (ORF 160a) and two on chromosome 3 (ORF 160b,c); the only other repeated genes are in the termini. Coding regions of some genes are shorter than their homologues in other organisms.

Only 17 protein-coding genes contain spliceosomal introns (42–52 bp long; see Supplementary Information), all located in the 5′ region, as in yeast[14], many immediately after the initiator AUG. Eleven are in ribosomal protein genes, as in yeast where their splicing negatively regulates messenger RNA levels[14]. Like the even shorter pygmy introns of chlorarachnean nucleomorphs[12], they have standard GT/AG boundaries. Twelve transfer RNA genes have protein-spliced introns. The marked contrast between the effective elimination of non-coding DNA from cryptomonad nucleomorphs and the accumulation of vast amounts of non-coding DNA in coexisting cryptomonad nuclei indicates that nuclear non-coding DNA in general is functional and positively selected[5], and is not purely selfish or junk. Chromosomal A+T content varies, suggesting that there have been three different mutational and selective pressures on base composition: on the terminal repeats (45.5% G+C); on housekeeping genes with a very low G+C content (23%); and on transfer RNA genes and genes for plastid proteins with intermediate G+C content (35%).

The function of 219 of the 464 putative protein-coding genes is unknown, but 31 have convincing database matches—11 to cyanobacteria and 20 to eukaryotes. Retention of the latter, even in this exceptionally compacted genome, shows that they must have important functions in all eukaryotes. The other 245 genes have homologues of known function, mostly for chromosome reproduction or gene expression, with very few for cytoplasmic functions such as protein assembly and degradation, signal transduction/regulation, cell-cycle control and membrane transport (Fig. 3). Only one gene for metabolism (carotenoid synthesis) was found. There are 47 different genes for non-mRNAs (rRNA, tRNA, small nuclear RNA, small nucleolar RNA).

Three conclusions can be drawn about intergenome cooperation in these complex eukaryote/eukaryote chimaeras. First, most identified genes (> 250/302) are needed simply for self-perpetuation of the nucleomorph and its periplastid ribosomes. End products directly useful to the rest of the cell are encoded by only a few genes including 30 chloroplast proteins, 3 transporters (Sut1 for sulphate; Kea1 for potassium; Rli1, an ABC transporter), an anabolic enzyme (Ggt), and a few regulatory enzymes. Second, even fewer cryptomonad plastid proteins are encoded by the nucleomorph than by the plastid genome[15], so at least a thousand more[1] must be imported into the plastid across four membranes, not two membranes as for nucleomorph-encoded ones. Third, as certain well-conserved genes essential for nucleomorph functions are absent from the nucleomorph genome, these functions must be provided by nuclear genes and imported into the periplastid compartment.

It was not known previously that nuclear gene products not destined for the chloroplast were imported into the periplastid space[4]. As in mitochondria and chloroplasts, DNA polymerase genes are absent. DNA polymerases must be nuclear encoded and imported across the RER and periplastid membranes, and onwards into the nucleomorphs through their nuclear pores. As we identified so few genes for transporters, most must be encoded by nuclear genes or nucleomorph ORFs. As chromobiotes lost the nucleomorph, they are likely to have homologous transport proteins in their nuclei. If chromists are sisters of alveolate protozoa, sharing a photosynthetic common ancestor[1,4,16], some may be also present in Sporozoa, such as malaria parasites, and important for their periplastid membranes and as potential drug targets.

Under 10% of the genes encode end-product functions[5] that are useful to the rest of the cell. Originally three end-product functions

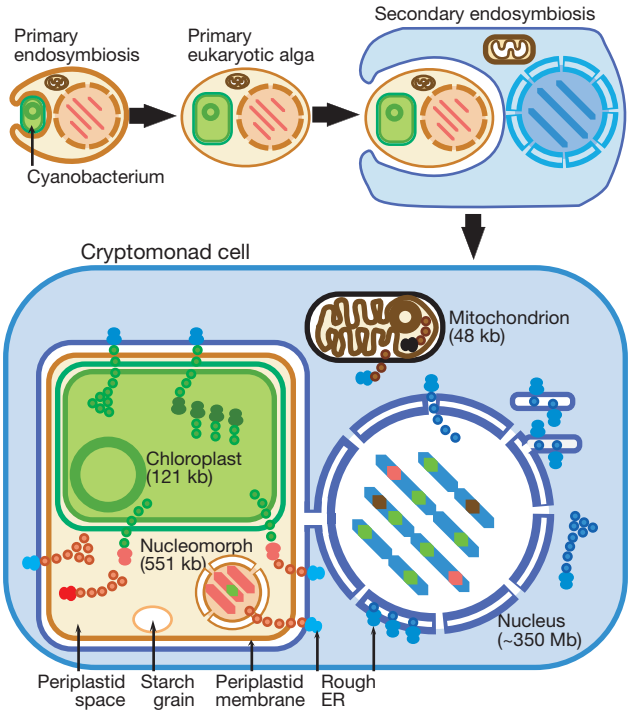


Figure 1 Secondary symbiogenetic origin and membrane topology of cryptomonads. After the primary endosymbiotic incorporation of a cyanobacterium to form the first chloroplast (green), many of its genes were transferred into the host nucleus. After this ancestral plant diversified to form green plants, red algae and glaucophytes, more complex algae were formed by independent secondary symbioses involving green or red algae, after which many or all plastid protein genes were transferred from the algal to the secondary host's nucleus. Shown here is the symbiosis of a red alga to form cryptomonads, where (as in all chromists) the food vacuole membrane fused with the RER. This fusion did not occur in alveolates, chlorarachneans or euglenoids. Former cyanobacterial genes now inserted in nucleomorph or nuclear chromosomes are shown in green, and former red algal genes now in the host nucleus in red. In cryptomonads, the chloroplast and nucleomorph (former red algal nucleus) are topologically in the periplastid space (starch- and ribosome-containing residual cytoplasm of the former red algal cell, yellow) in the periplastid membrane (former red algal plasma membrane), which is located in the lumen of the host's rough endoplasmic reticulum (RER). Chloroplast proteins are coded by three genomes (chloroplast, nucleomorph and nucleus) and mitochondrial proteins by two genomes (mitochondrion and nucleus). Nucleomorph and periplastid proteins are coded by two genomes (nucleus and nucleomorph). Coloured dots indicate protein translocation pathways in the periplastid complex: nuclear- or nucleomorph-encoded proteins targeted to the chloroplast are green; nuclear-encoded proteins imported into the periplastid space, and both nuclear- and nucleomorph-encoded proteins imported into the nucleomorph, are red. Mitochondrial proteins (brown) are encoded by both nuclear and mitochondrial genomes.

© 2001 Macmillan Magazines Ltd

Figure 2 The cryptomonad nucleomorph chromosomes showing gene locations. Colours show the functional categories of the genes specified in Fig. 3. The number of overlapping nucleotides between adjacent genes is indicated beside the brackets. Each chromosome is displayed as if broken into two fragments near its midpoint. Labels on the left of each chromosome indicate genes transcribed towards its left end (top left); labels on the right show those transcribed towards the right end (bottom right).

© 2001 Macmillan Magazines Ltd

were envisaged: plastid proteins not made in the chloroplast itself or in the host cytosol; periplastid space functions in starch metabolism; and periplastid membrane transporters mediating metabolite exchange between host and symbiont. We identified 30 genes encoding proteins for chloroplast function, including two known proteins[11], but very few for periplastid functions.

Sequence comparisons show that two nucleomorph genes for plastid proteins are eukaryotic inventions (Iap100, Met), one (CbbX) is α-proteobacterial[17], and the others are cyanobacterial. Only two encode molecules directly needed for photosynthesis: an electron transfer molecule (rubredoxin)[11] and a carotene-binding protein (Hlip). Some genes are needed for chloroplast division (FtsZ[11]) or gene expression (ribosomal protein Rps15, products for DNA and RNA metabolism, and a factor for RuBisCo expression, CbbX). Others encode components of the chloroplast protein-import machinery (Iap100, Tic22), translocation into the thylakoid lumen (Tha4, SecE), chaperonins (Cpn60, Hcf136) and a protease (ClpP).

All are essential for chloroplast function, explaining why the nucleomorph has persisted for hundreds of millions of years. Compared with their cyanobacterial homologues, the nucleomorph-encoded plastid proteins have amino-terminal extensions (transit peptides; see Supplementary Information) to specify import into the chloroplast, although only those of Hcf136, Cpn60, Tic22 and GyrB were recognized by the search tool ChloroP (http://www.cbs.dtu.dk/services/ChloroP/). Nevertheless, *in vitro* the N-terminal extension of rubredoxin acts as a transit peptide with pea or cryptomonad plastids and is removed after import[18].

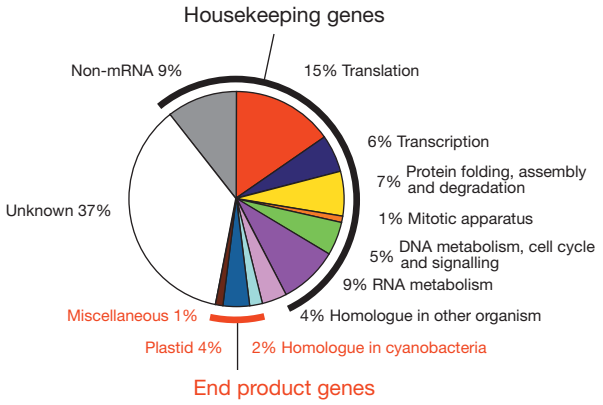


Figure 3 The fraction of genes in each functional category and their classification as end-product or genetic housekeeping genes. The genes in each category are listed in the Supplementary Information.

Acquiring solar-powered carbohydrate synthesis is the main advantage of enslaving a photosynthetic cell. As glucose produced by degrading periplastid starch is used mainly in the host cytosol, starch metabolism must be responsive to its needs. If the phosphatidylinositol-4-OH kinase (PI(4)K) is associated with the periplastid membrane, it may mediate such signalling through the second messenger inositol trisphosphate. In response, nucleomorph-encoded GTP-binding proteins (GbIp, GTPbp) may regulate periplastid starch metabolism and protein synthesis, as may a protein kinase similar to glycogen synthase kinase; this AMP-activated kinase (AakB) might control starch degradation (like its homologue in animal glycogen degradation). The nucleomorph has four other kinase genes (gs, mps1, snf1, snf2). Protein phosphatase, Pp1, may function complementarily in starch regulation and/or kinase-regulated cell-cycle controls. The presence of a geranylgeranyl-transferase (Ggt) lacking a transit peptide indicates that at least one intermediate stage in carotenoid synthesis may occur in the periplastid space. As plastids typically make carotenoids using isoprenoid precursors from the host, intermediates presumably traverse the periplastid space.

The red algal symbiont lost mitochondria, peroxisomes, lysosomes, Golgi membranes and most metabolic enzymes. Possibly it also lost all ER lumenal proteins and protein glycosylation, perhaps facilitating an unprecedented simplification of eukaryotic membrane-protein targeting. This red algal vestige may differ from true eukaryotic and bacterial cells by lacking co-translational protein insertion into membranes by the signal recognition particle (SRP): the nucleomorph outer membrane, phylogenetically equivalent to RER[8], uniquely lacks obvious surface ribosomes[13,19], and we found no 7S SRP RNA or other genes for SRPs. Perhaps nucleomorph envelope and periplastid membrane proteins are inserted post-translationally, as in mitochondria and plastids[1] (which analogously both lost the ancestral eubacterial SRP RNA), using the nucleomorph-encoded Hsp70 chaperone.

Genes for DNA polymerase are absent, but there are genes for its sliding clamp (Pcna), a replication co-factor (RfC) for primer extension, and a RecA-like protein (Rad51). Replicon origins might lie in the terminal repeats proximal to the 18S rRNA genes, where there is more non-coding DNA than elsewhere; if there are no others, each chromosome will consist of just two 80–100-kilobase (kb) replicons. As eukaryotic replicons average 100 kb (ref. 20), no origins need lie in unique chromosome regions where intergenic spacers are almost all shorter than the minimal 250 bases needed for origins.

Genes for three core histones (H2b, H3, H4) are present. H2b is

Table 1 Codon usage, introns and chromosomal location of tRNA genes

Codon	Chr*	%	Codon	Chr*	%	Codon	Chr*	%	Codon	Chr*	%
TTC F	2†	21.4	TCC S		5.6	TAC Y	1†	20.2	TGC C	3†	24.6
TTT F		78.6	TCT S	2†	34.1	TAT Y		79.8	TGT C	2†	75.4
TTA L	3†	50.8	TCA S		34.1	TAA Z		60.0	TGA Z		20.0
TTG L	2†	17.6	TCG S	1	5.6	TAG Z		20.0	TGG W	3†	100
CTC L		1.9	CCC P		12.6	CAC H	2	18.4	CGC R		0.8
CTT L	3	14.9	CCT P	3	32.3	CAT H		81.6	CGT R	3	2.4
CTA L	1	12.2	CCA P		48.3	CAA Q	1	78.2	CGA R	2	8.1
CTG L		2.7	CCG P	1	6.9	CAG Q	2	21.8	CGG R		2.4
ATC I		12.9	ACC T		10.6	AAC N	3	23.6	AGC S	2	6.0
ATT I	1	48.6	ACT T	2	34.1	AAT N		76.4	AGT S		14.5
ATA I	2	38.5	ACA T	3	49.2	AAA K	2	87.1	AGA R	1†	77.4
ATG M_i	3	100	ACG T	3	6.2	AAG K	2	12.9	AGG R	3†	8.9
ATG M_e	3†	100									
GTC V		7.5	GCC A		9.1	GAC D	1	18.6	GGC G	3	9.2
GTT V	2†	48.4	GCT A		32.6	GAT D		81.4	GGT G		36.2
GTA V	1	34.4	GCA A	3	50.8	GAA E		84.1	GGA G	3	48.5
GTG V		9.7	GCG A		7.6	GAG E		15.9	GGG G		6.1

M_e, elongator methionine; M_i, initiatior methionine.
* Chromosome on which tRNA gene is located.
† tRNA gene contains non-spliceosomal intron.

© 2001 Macmillan Magazines Ltd

exceptionally divergent with its normally acetylated basic N-terminal tail mostly deleted, so nucleosomes may be simplified compared with those of typical nuclei. Histones probably undergo acetylation, as genes for a histone acetyltransferase (Hat) and deacetylase (Hda) are found, and all H3 and H4 acetylation sites are conserved. H3 and H4 genes are adjacent, divergently transcribed and relatively well-conserved, but have evolved more rapidly than nuclear homologues (like most nucleomorph genes). Although H2a and H1 are undetectable, neither is likely to be dispensable, as a string of nucleosomes, unlike a 30-nm filament requiring H1, would be too long for segregation without them.

One gene each for α-, β-, and γ-tubulin is present[11], and in addition to γ-tubulin three other centrosomal proteins[21] are nucleomorph-encoded—Ranbpm[11]; Hsp82 (in the cytosolic Hsp90 family) and Hsp 70 (cytosolic family). As centrosome activation involves the cyclinB–cdc2 complex[22] (both nucleomorph encoded), mitotic control is present. Nuclear localization signals on Hsp82 and Hsp70 suggest that the centrosome is inside the nucleomorph, consistent with the intact envelope during division[19] and the intranuclear red algal spindles. Presence of a centromere-specific H3-like histone (Cenp-A) suggests that nucleomorph chromosomes have centromeres, confirming that nucleomorphs have a relict mitotic apparatus[11] despite the lack of ultrastructural evidence for a spindle[19].

We cannot be certain of centromere positions. Only the terminal repeats have significant non-coding DNA regions repeated on all chromosomes, but we doubt that they are centromeric because two per chromosome would upset segregation. Assuming nucleomorph centromere sizes to be similar to *Saccharomyces cerevisiae*[23], chromosome 1 has two non-genic regions long enough to be centromeric (520 nucleotides between ORFs 147 and 471; 540 nucleotides between ORFs 244 and 446); chromosome 2 has one (~200 nucleotides between ORFs 80a and 231); chromosome 3 has two (374 nucleotides between ORFs 141 and 180; ~300 nucleotides between ORF 62 and Ebi). As these candidate centromere regions are near chromosome midpoints and other intergenic regions seem too short, all three chromosomes may be metacentric.

Eukaryotic chromosomes are linear arrays of looped domains attached to a nuclear skeleton by their replicon origins[24]. Each nucleomorph chromosome may constitute a single loop domain, being similar in size. If nucleomorph histones compact the DNA into typical 30-nm chromatin threads, the longest chromosome would be 1.5 μm in the absence of higher-order condensation; if metacentric, each chromosome arm would be 0.75 μm. As nucleomorphs are about 1.5-μm across and longer when dividing[19], extra folding would be unneccessary for mitotic segregation, unlike for other eukaryotic chromosomes. Compacted chromosomes cannot be seen during nucleomorph division[19]. Analogous folding constraints may explain why cryptomonad and chlorarachnean nucleomorphs retained three short chromosomes[3,9,12] instead of aggregating them into a larger one, which would be feasible only with higher-order mitotic compaction.

Genes for key cell-cycle control proteins include a replication-licensing protein (Mcm2), which mediates the G1/S-phase transition[22], and a cyclin-dependent Cdc2 kinase and its cyclin B, which are involved in the G2/M-phase checkpoint. The persistence of a nucleomorph-specific cell-cycle kinase system, even in this simplified cell vestige, emphasizes that it remains conceptually equivalent to a cell, despite its long integration and loss of metabolism and most genes. The nucleomorph encodes ubiquitin (fused to two ribosomal proteins), the ubiquitin-fusion-degradation enzyme (Ufd), and three E2 enzyme subunits (Ubc2, Ubc4, UbcE2). The cell-cycle regulatory role of proteasomes (degrading cyclin B to ensure exit from mitosis, and also controlling exit from S phase[25]) may be the key reason why such complex multiprotein assemblies as proteasomes and their associated ubiquitin pathway are conserved in the periplastid space, otherwise so reduced in normal cytosolic functions. Supporting this is a gene for an AAA-ATPase (Cdc48; plus a partially related gene *cdc48a*) implicated in ubiquitin-dependent mitotic events, including membrane fusion[26]. There are 21 nucleomorph genes for different subunits of the 20S core proteasome and its 19S cap. Two 20S proteasome subunits (α4 and β2) and ten 19S cap subunits are absent, and must be imported like all 11S activator subunits.

Elements of the nuclear pore-complex export/import machinery, the key nucleolar structural protein fibrillarin (Nop1), and elaborate RNA processing machinery confirm that the nucleomorph is a miniaturized nucleus. The surmise[13] that nucleomorph envelope pores are genuine nuclear pores is confirmed by well-conserved importin genes (*impA*; *imb1*) and a protein interacting with nuclear pores and needed for transport (CRM). Neither transcription nor RNA processing was radically simplified by symbiogenesis. All three RNA polymerases and several transcription factors are nucleomorph encoded, including TfIID for promoter recognition, TfIIB for binding polymerase II, and two components (Rad3; Rad25) of TfIIH. Nucleomorph messengers are probably capped and polyadenylated, for mRNA-capping enzyme (Mce), cap-binding protein (Cbp), and poly(A)-binding protein (Pab) genes exist. Many spliceosomal splicing components (including U6 snRNA and snRNPE, as in chlorarachneans[12]) and some for tRNA intron removal are present, as are 5 RNAs and 17 proteins of the nucleolar snoRNP machinery for processing rRNA (cleavage, methylation and pseudouridylation). The gene for pseudouridylate synthase is of the eukaryotic type (*cbf5*) with ancillary centromere–microtubule binding properties, but lacks several normally conserved features. The thesis that snoRNAs are ancestral for all cells but were lost in eubacteria by 'streamlining'[27] is rendered implausible by their persistence in this ultra-streamlined genome.

Thirty-seven genes encode large subunit periplastid ribosomal proteins and 28 small subunit ones, so about 14 are imported or coded by unidentified ORFs. The nucleomorph encodes 37 tRNAs (12 with introns; Table 1) and standard protein synthesis initiation and elongation factors, but only one amino-acyl-tRNA synthetase (seryl). As we found no gene for glutamyl-tRNA, it must be imported unless its codons are recognized by another tRNA with 'super-wobble'.

Cryptomonad and chlorarachnean nucleomorphs are natural experiments in genome miniaturization and cell simplification that can test basic ideas about genome and cell functions. If domain organization, folding, and mitotic segregation of nucleomorph chromosomes are indeed exceptionally simplified, they could become important models for understanding the more complex chromosomes of typical nuclei. □

Methods

Nucleomorph DNA

As nucleomorph DNA is only about 0.1% of cellular DNA, we could not obtain pure enough samples or sufficient amounts of it for a single mechanically fragmented library. We prepared several libraries slightly contaminated by DNA from the other three genomes: random libraries from the nucleomorph DNA band on CsCl density gradients[28] (with major mitochondrial DNA contamination) and others enriched in one individual nucleomorph chromosome extracted from bands excised from pulsed field gels[9]. DNA was partially digested by *Sau*3AI and cloned into phage λ as described[11], or completely digested with one of *Eco*RI, *Hind*III, *Pst*I, *Bgl*II, *Xba*I, *Cla*I, *Spe*I, *Bcl*I or *Bam*HI and cloned into plasmids (pUC18).

Sequencing and gene identification

Sequencing, editing and contig assembly were done as described[11]. Clones were sequenced fully on both strands by primer walking, and gaps between contigs filled by polymerase chain reaction of genomic DNA. Preliminary gene identification and analysis used MAGPIE[29]. tRNA and snoRNA genes were detected using the search tools tRNAscan (http://www.genetics.wustl.edu/eddy/ tRNAscan-SE) and snoscan[30]. Further gene analyses used a Perl script to identify all possible ORFs and search them against GenBank (BLAST results available at http:// reith.imb.nrc.ca/nucleomorph/nucleomorph.html). Spliceosomal introns were first found manually, and then using a Perl script that searched for conserved splice junctions and extension of the ORF by removing the putative intron.

© 2001 Macmillan Magazines Ltd

Received 5 September 2000; accepted 14 February 2001.

1. Cavalier-Smith, T. Membrane heredity and early chloroplast evolution. *Trends Plant Sci.* **5,** 174–182 (2000).
2. Douglas, S. E., Murphy, C. A., Spencer, D. F. & Gray, M. W. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. *Nature* **350,** 148–151 (1991).
3. Maier, U.-G., Douglas, S. & Cavalier-Smith, T. The nucleomorph genomes of cryptophytes and chlorarachniophytes. *Protist* **151,** 103–109 (2000).
4. Cavalier-Smith, T. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Euk. Microbiol.* **46,** 347–366 (1999).
5. Cavalier-Smith, T. & Beaton, M. J. The skeletal function of non-genic nuclear DNA: new evidence from ancient cell chimaeras. *Genetica* **106,** 3–13 (1999).
6. Cavalier-Smith, T. A revised six-kingdom system of life. *Biol. Rev.* **73,** 203–266 (1998).
7. Moreira, D., Le Guyader, H. & Philippe, H. The origin of red algae: implications for the evolution of chloroplasts. *Nature* **405,** 69–72 (2000).
8. Cavalier-Smith, T. in *Progress in Phycological Research* Vol. 4 (eds Round, F. E. & Chapman, D. J.) 309–347 (Biopress, Bristol, 1986).
9. Eschbach, S., Hofmann, C. J. B., Maier, U.-G., Sitte, P. & Hansmann, P. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga, *Pyrenomonas salina*. *Nucleic Acids Res.* **19,** 1779–1781 (1991).
10. Whatley, J. M., John, P. & Whatley, F. R. From extracellular to intracellular: the establishment of mitochondria and chloroplasts. *Proc. Roy. Soc. Lond. B.* **204,** 165–187 (1979).
11. Zauner, S. *et al.* Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. *Proc. Natl Acad. Sci. USA* **97,** 200–205 (2000).
12. Gilson, P. R. & McFadden, G. I. The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. *Proc. Natl Acad. Sci. USA* **93,** 7737–7742 (1996).
13. Ludwig, M. & Gibbs, S. P. Are the nucleomorphs of cryptomonads and *Chlorarachnion* the vestigial nuclei of eukaryotic endosymbionts? *Ann. NY Acad. Sci.* **503,** 198–211 (1987).
14. Spingola, M., Grate, L., Haussler, D. & Ares, M. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5,** 221–234 (1999).
15. Douglas, S. E. & Penny, S. L. The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.* **48,** 236–244 (1999).
16. McFadden, G. I. Mergers and acquisitions: malaria and the great chloroplast heist. *Genome Biol.* **1,** 1026.1–1026.4 (2000).
17. Maier, U. G., Fraunholz, M., Zauner, S., Penny, S. & Douglas, S. A nucleomorph-encoded CbbX and the phylogeny of RuBisCo regulators. *Mol. Biol. Evol.* **17,** 576–583 (2000).
18. Wastl, J. & Maier, U. -G. Transport of proteins into cryptomonads complex plastids. *J. Biol. Chem.* **275,** 23194–23198 (2000).
19. Meyer, S. The taxonomic implications of the ultrastructure and cell division of a stigma-containing *Chroomonas* sp. (Cryptophyceae) from Rocky Bay, Natal, South Africa. *S. Afr. J. Bot.* **53,** 129–139 (1987).
20. Berezney, R., Dubey, D. D. & Huberman, J. A. Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma* **108,** 471–484 (2000).
21. Lange, B. H. M., Bachi, A., Wilm, M. & Gonzalez, C. Hsp90 is a core centrosomal component and is required at different stages in the centrosome cycle in *Drosophila* and vertebrates. *EMBO J.* **19,** 1252–1262 (2000).
22. De Souza, C. P., Ellem, K. A. & Gabrielli, B. G. Centrosomal and cytoplasmic Cdc2/cyclin B1 activation precedes nuclear mitotic events. *Exp. Cell Res.* **257,** 11–21 (2000).
23. Hegemann, J. H. & Fleig, U. N. The centromere of budding yeast. *BioEssays* **15,** 451–460 (1993).
24. Paul, A. L. & Ferl, R. J. Higher-order chromatin structure: looping long molecules. *Plant Mol. Biol.* **41,** 713–720 (1999).
25. Kawahara, H. *et al.* Inhibiting proteasome activity causes overreplication of DNA and blocks entry into mitosis in sea urchin embryos. *J. Cell Sci.* **113,** 2659–2670 (2000).
26. Meyer, H. H., Shorter, J. G., Seemann, J., Pappin, D. & Warren, G. A complex of mammalian ufd1 and npl4 links the AAA-ATPase, p97, to ubiquitin and nuclear transport pathways. *EMBO J.* **19,** 2181–2192 (2000).
27. Poole, A., Jeffares, D. & Penny, D. Early evolution: prokaryotes, the new kids on the block. *BioEssays* **21,** 880–889 (1999).
28. Douglas, S. E. Physical mapping of the plastid genome from the chlorophyll c-containing alga, *Cryptomonas* ϕ. *Curr. Gen.* **14,** 591–598 (1988).
29. Gaasterland, T. & Sensen, C. W. Fully automated genome analysis that reflects user needs and preferences: A detailed introduction to the MAGPIE system architecture. *Biochimie* **78,** 302–310 (1996).
30. Lowe, T. M. & Eddy, S. R. A computational screen for methylation guide snoRNAs in yeast. *Science* **283,** 1168–1171 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We are grateful to M. Johannsen for technical assistance; P. Gordon for assistance with MAGPIE; T. Lowe for helping us identify tRNAs and snoRNAs; and R. Redfield for preparing Fig. 1 and valuable comments on the manuscript. Our research was supported by the Natural Sciences and Engineering Council (NSERC), Canada, the Deutsche Forshungsgemeinschaft, Germany, and the NERC (UK). We thank the Canadian Institute for Advanced Research and NERC for fellowships for T.C.-S. and NSERC for a fellowship for M.J.B.

Correspondence and requests for materials should be addressed to T.C.-S. (e-mail: tom.cavalier-smith@zoo.ox.ac.uk). Sequences have been deposited in GenBank under accession numbers AF165818, AJ010592 and AF083031.

Nonlinear effects of large-scale climatic variability on wild and domestic herbivores

Atle Mysterud*, Nils Chr. Stenseth*, Nigel G. Yoccoz*†, Rolf Langvatn‡ & Geir Steinheim§

* *Department of Biology, Division of Zoology, University of Oslo, PO Box 1050 Blindern, N-0316 Oslo, Norway*
† *Department of Arctic Ecology, Norwegian Institute for Nature Research (NINA), Polar Environmental Centre, N-9296 Tromsø, Norway*
‡ *University Courses on Svalbard (UNIS), N-9170 Longyearbyen, Spitsbergen, Norway*
§ *Department of Animal Science, Agricultural University of Norway, PO Box 5025, N-1432 Ås, Norway*

Large-scale climatic fluctuations, such as the North Atlantic Oscillation (NAO)[1,2], have been shown to affect many ecological processes[3–6]. Such effects have been typically assumed to be linear. Only one study has reported a nonlinear relation[7]; however, that nonlinear relation was monotonic (that is, no reversal). Here we show that there is a strong nonlinear and non-monotonic (that is, reversed) effect of the NAO on body weight during the subsequent autumn for 23,838 individual wild red deer (*Cervus elaphus*) and 139,485 individual domestic sheep (*Ovis aries*) sampled over several decades on the west coast of Norway. These relationships are, at least in part, explained by comparable nonlinear and non-monotonic relations between the NAO and local climatic variables (temperature, precipitation and snow depth). The similar patterns observed for red deer and sheep, the latter of which live indoors during winter and so experience a stable energy supply in winter, suggest that the (winter) climatic variability (for which the index is a proxy) must influence the summer foraging conditions directly or indirectly.

Local weather patterns[8] and large-scale climatic variability[6,9–11] affect demography and population dynamics of temperate ungulates. The NAO[1] is a large-scale fluctuation in atmospheric mass between the subtropical North Atlantic region (centred on the Azores) and the subpolar North Atlantic region (centred on Iceland)[12]. The NAO is positively correlated with temperature and precipitation at the west coast of Norway during winter[1]. Because temperatures in this region often are around 0 °C during winter, precipitation (as rain) is positively correlated and snow depth is negatively correlated with the NAO at low elevation (below 400 m), and precipitation (as snow) is positively correlated with the NAO at high elevation[11]. Similar differential impacts of global climatic warming at low-altitude and high-altitude habitats have been reported from the Rocky Mountains[13].

Using wild red deer (*Cervus elaphus*) and domestic sheep (*Ovis aries*) lambs as examples of wild and domestic herbivores, we have explored comparatively the effect of the NAO (being a proxy for the interannual variation in winter temperature and precipitation along the west coast of Norway[1,11]) on individual performance (measured as body weight). As domestic sheep in Norway are free-ranging only from May/June to September/October, we can separate, through a comparison of the two species, the direct effects (costs of thermoregulation and movement in snow) and indirect effects (delayed snow melt can affect foraging conditions during summer) of winter climate.

In our study region, sheep and red deer in Norway are sympatric and have very similar diet composition and habitat use[14]. During winter, however, sheep are fed indoors, and as a result experience a stable interannual energy supply during that season. Thus, any relation between growth of lambs (during summer) and winter

© 2001 Macmillan Magazines Ltd

climate must operate through indirect effects of winter climate on forage quantity and quality during summer.

A main positive effect of the NAO was found for both male and female red deer, with a plateau being reached for high NAO values (Fig. 1, Table 1). There may be direct negative effects of severe winter conditions owing to increased energetic costs of thermoregulation and locomotion; snow also reduces access to high-quality forage[15]. This may lead to increased loss of body mass[16], retarded growth *in utero* and a delayed birth date[17], which later may affect survival and fecundity of the cohort[9,18]. High temperatures and less snow at low altitude may provide favourable winter conditions for red deer, as they descend to low elevation during the winter[19].

The decreasing proportion of male calves born after winters with a decreasing NAO index is most probably due to fetal mortality *in utero* and/or perinatally[11], and thus is evidence for some direct effect of winter climate. There is also a clear effect of the winter NAO index on autumn body mass of domestic sheep (Fig. 2, Table 2). Note that if there is no control for local density dependence (as in refs 6 and 9), the main positive effect of the NAO on red deer body weight will not be observed. The general positive relation between red deer body weight and the state of the NAO is reversed for the lower values of the NAO (Fig. 1), producing a nonlinear and non-monotonic response. This is the first study, to our knowledge, to report such a non-monotonic relation between the ecological performance of populations and large-scale climatic variability. The non-monotonic response of large herbivore populations to large-scale climatic variation is most probably related to a similar nonlinear and non-monotonic relationship between the NAO and local climatic variables on the west coast of Norway (Fig. 3).

As red deer and sheep showed the same response to winter climate, it is likely that it is the indirect effect of winter climate that produces the nonlinear effect. Forage quality of grasses and

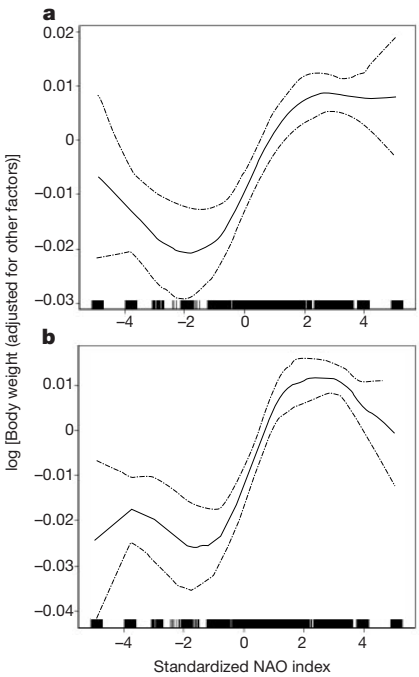


Figure 1 Relationship between autumn body weight of adult (≥ 1 yr) red deer and the North Atlantic Oscillation, an index of large-scale winter climatic variability. **a**, Female; **b**, male. Dashed lines are 95% point-wise confidence intervals. Tick marks show the locations of the observations on that variable[29].

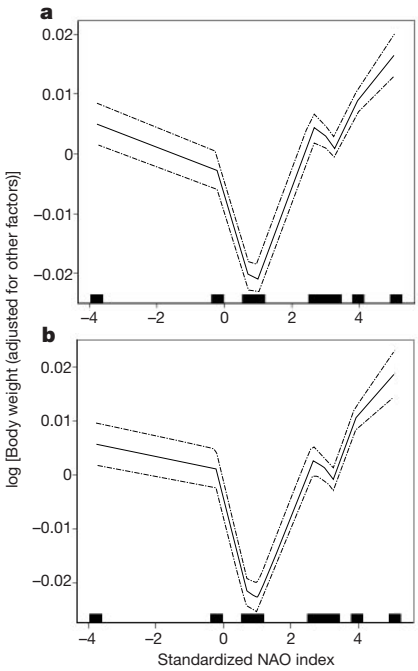


Figure 2 Relationship between autumn body weight of domestic sheep lambs, which are free-ranging only during summer, and the North Atlantic Oscillation, an index of large-scale winter climatic variability. **a**, Female; **b**, male. Dashed lines are 95% point-wise confidence intervals. Tick marks show the locations of the observations on that variable[29].

Table 1 Results from fitting general additive models on variation in autumn body weight of wild red deer

	d.f.	$d.f._{spline}$	*F*	*P*
Females				
NAO	1	3 versus 4	4.790	0.028
Age	1	4 versus 5	185.900	< 0.001
Density	1		3.282	< 0.001
Date of shooting	1	1 versus 2	16.701	< 0.001
Proportion of high-altitude habitat	1	4 versus 5	8.820	0.003
Degree of latitude	1	4 versus 5	6.310	0.012
Distance from the coast	1	4 versus 5	9.120	0.003
Population	4		20.290	< 0.001
Males				
NAO	1	4 versus 5	7.270	0.007
Age	1	4 versus 5	66.600	< 0.001
Density	1	2 versus 3	4.110	0.043
Date of shooting	1	4 versus 5	5.130	0.024
Proportion of high-altitude habitat	1	4 versus 5	21.470	< 0.001
Degree of latitude	1	4 versus 5	6.280	0.012
Distance from the coast	1	4 versus 5	9.570	0.002
Population	4		32.392	< 0.001

Data were from 9,175 female and 14,323 male wild red deer during the period 1965–1998.

Table 2 Results from fitting general additive models on variation in autumn body weight of sheep

	d.f.	$d.f._{spline}$	*F*	*P*
Females				
NAO	1	4 versus 5	100.66	< 0.001
Breed	1		2,264.400	< 0.001
Age of lamb	1	4 versus 5	23.204	< 0.001
Age of ewe	1	4 versus 5	60.826	< 0.001
Litter size	1	3 versus 4	5.120	0.039
Density	1	4 versus 5	49.022	< 0.001
Municipality	13		6.179	< 0.001
Males				
NAO	1	4 versus 5	118.000	< 0.001
Breed	1		1,158.100	< 0.001
Age of lamb	1	4 versus 5	11.393	< 0.001
Age of ewe	1	4 versus 5	60.654	< 0.001
Litter size	1	2 versus 3	32.222	< 0.001
Density	1	4 versus 5	57.137	< 0.001
Municipality	13		6.836	< 0.001

Data were from 71,812 female and 67,673 male sheep lambs during the period 1989–1998.

© 2001 Macmillan Magazines Ltd

herbs declines with age (time after emergence)[20]. Hence, it is commonly assumed that winter climate may also indirectly affect foraging conditions during summer, as deep snow may lead to an extended period of snow melt and hence a prolonged period of access to newly emergent high-quality forage[6,10,19]. Indeed, there is an increase in variability in timing of flowering with increasing NAO index[6]. The prolonged period with access to newly emergent, high-quality forage is favourable to both red deer and sheep. Wheat quality in the UK has been positively correlated with the NAO[21], suggesting that there may be an effect of the NAO on plant quality over and above its effect on time of emergence. Even slight changes in forage quality may substantially affect body growth of ungulates[22]. Our results therefore provide unambiguous support for an indirect effect of winter climate on summer foraging conditions.

A trend of increasingly warm and wet winters (as is typical for years with high NAO values) may therefore favour large herbivorous ungulates along the west coast of Norway through two separate mechanisms: first, less snow in the low-elevation wintering areas will decrease energetic costs of thermoregulation and movement, and increase access to forage in the field layer during winter[10]; and second, more snow in the high-elevation summer areas will lead to a prolonged period of access to high-quality forage during summer.

As body weight in ungulates is closely related to survival[10], reproduction in females[23,24] and fighting ability in males[17], the population dynamical implications of the variability in the NAO on large herbivores may be pronounced. Our result emphasizing the nonlinear and non-monotonic ecological response to large-scale climatic variability clearly demonstrates the difficulty in predicting the ecological impact of large-scale climatic fluctuations, and emphasizes the importance of routinely assessing nonlinearity in studies on global change. □

Methods

Data

The data used in this paper derive from various municipalities in six Norwegian counties: Rogaland (n_{deer} = 6, n_{sheep} = 3); Hordaland (n_{deer} = 23, n_{sheep} = 3); Sogn and Fjordane (n_{deer} = 22, n_{sheep} = 4); Møre and Romsdal (n_{deer} = 30, n_{sheep} = 3); Sør-Trøndelag (n_{deer} = 17, n_{sheep} = 6); and Nord-Trøndelag ($n_{red\ deer}$ = 7). Data on dressed weight of 23,838 male and female red deer were sampled during the annual autumn harvest 1965–1998 (refs 9, 24). Data on covariates such as age (estimated using annuli in the cementum of the first incisor[25]), an index for local density (total harvest in a municipality divided by the area of red deer habitat[11,25]), date of culling[9,25], distance from the coast[25,26], degree of latitude[26] and proportion of high-altitude habitat[25] were retrieved from earlier studies.

Data on body weight, sex, breed (Dala and Spel), age of ewe (years), age of lamb (days), litter size and location (municipality) of 139,485 lambs[27] from 1989 to 1998 were retrieved from the Norwegian sheep recording system. These are lambs that have been free-ranging all summer, and that are weighed shortly after they are gathered from the summer pastures. The ages of ewes and lambs were known precisely, because year and date of birth had been recorded. The sheep data were extracted from the municipalities with the most red deer data. Stocking densities of sheep on pastures was calculated as total number of sheep (for each year in each muncipality; known from the official agricultural statistics, Statens Kornforretning) divided by total area of the municipalities.

As an index of global winter climate we used the NAO[1], which is a large-scale fluctuation

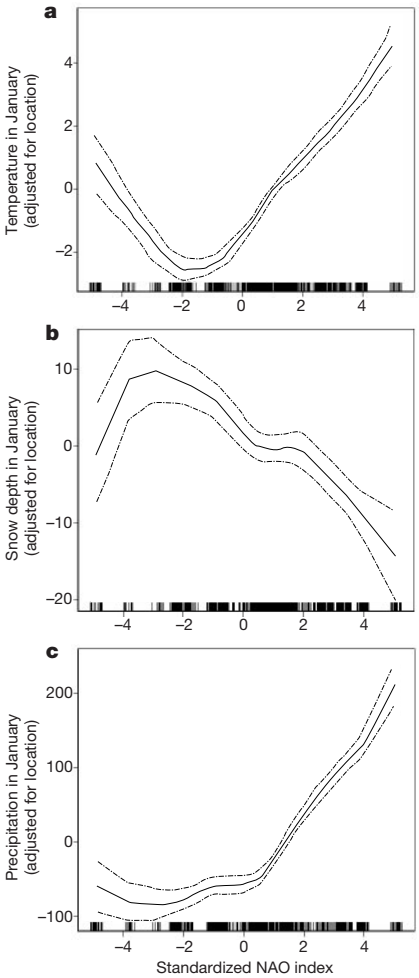


Figure 3 Relationship between the local climatic variables at the west coast of Norway and the North Atlantic Oscillation, an index of large-scale winter climatic variability. **a**, Temperature (d.f. = 1, d.f.$_{splines}$ = 4 versus 5; F = 8.100, P = 0.005). **b**, Snow depth (d.f. = 1, d.f.$_{splines}$ = 4 versus 5; F = 9.758, P = 0.002), **c**, Precipitation in January (d.f. = 1, d.f.$_{splines}$ = 4 versus 5; F = 16.200, $P < 0.001$). Dashed lines are 95% point-wise confidence intervals. Tick marks show the locations of the observations on that variable[29].

© 2001 Macmillan Magazines Ltd

in atmospheric mass between the subtropical North Atlantic region (centred on the Azores) and the subpolar North Atlantic region (centred on Iceland)[12]. The NAO index used was Hurrell's winter index[1], which is based on the difference of normalized sea level pressures between Lisbon, Portugal and Stykkisholmur/Reykjavik, Iceland for the months December to March (http://goldhill.cgd.ucar.edu/cas/climind/nao_winter.html). Similar results to those reported in the main text may also be obtained by using indexes based on either the Azores (http://goldhill.cgd.ucar.edu/cas/climind/nao_monthly.html) or Gibraltar (http://www.met.rdg.ac.uk/cag/NAO/) as a southern station. Data on local climatic variability (temperature, snow depth and precipitation) from 1964 to 1998 on and 39 meteorological stations (restricted to stations below 250 m above sea level, as there is interaction between snow depth and altitude[11]) were obtained from The Norwegian Meteorological Institute, Oslo.

Statistical analyses

We used general additive models (GAM) with smoothing splines[28,29] that allow nonlinear functions of the independent variables to be chosen by the procedures, using the function GAM in the statistical package S-Plus[29]. Splines are piecewise polynomials being smoothed (twice differentiable) in the spline knots (the control points of the splines). The complexity of the curve (that is, the number of degrees of freedom; $d.f._{spline}$) associated with the smoothing spline was selected by repeated fitting of the GAM with varying $d.f._{spline}$ (1–5) for one variable, while holding $d.f._{spline}$ of the other variables constant. The fits of the different models were then tested in an analysis of variance setting[29], and the selected $d.f._{spline}$ was used in the graphical presentation (the maximum significant $d.f._{spline}$ used). The weights of red deer and sheep were log-transformed to stabilize the variance.

For red deer, we used separate models for males and females (all ≥ 1 yr), as the response was expected to be variable among sex classes[30], and because interactions cannot be included in GAM[28]. In both models for red deer, we controlled for the curvilinear effect of age and proportion of high-altitude habitat[25], the close to linear positive effect of distance from the coast[26] and degree of latitude[26], and the negative effect of date of culling and of local density[25], and the factorial variable "population"[25].

For sheep, we used separate models for males and females and included, in addition to the NAO, the factorial variables breed and municipality, the close to linear negative effect of litter size and density, the close to linear positive effect of age of lamb, and the curvilinear effect of age of ewe[27].

The lowest autumn body weights of red deer were recorded at a value of about −2 of the NAO index. Four years with a particularly low NAO value (1969, 1977, 1979 and 1996)[1] were responsible for producing this result. As sheep data (1989–1998) only contained a particularly low NAO value for 1996, the slightly different estimate for the NAO producing the lowest body weight (around NAO = 1) may be due to absence of data around a NAO index of −2. Detrending the NAO time series (by adding a smooth spline fit for 'year' to the models) did not alter the conclusion reported in the main text.

Because there is variation due to altitude, latitude and longitude[11], we included meteorological station as a factor variable in addition to the NAO for local climatic variability.

Received 29 November 2000; accepted 22 January 2001.

1. Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269,** 676–679 (1995).
2. Stephenson, D. B., Pavan, V. & Bojariu, R. Is the North Atlantic Oscillation a random walk? *Int. J. Clim.* **20,** 1–18 (2000).
3. Forchhammer, M. C., Post, E. S. & Stenseth, N. C. Breeding phenology and climate. *Nature* **391,** 29–30 (1998).
4. Crick, H. Q. P. & Sparks, T. H. Climate change related to egg-laying trends. *Nature* **399,** 423–424 (1999).
5. Hughes, L. Biological consequences of global warming: is the signal already apparent? *Trends Ecol. Evol.* **15,** 56–61 (2000).
6. Post, E. & Stenseth, N. C. Climatic variability, plant phenology, and northern ungulates. *Ecology* **80,** 1322–1339 (1999).
7. Sæther, B.-E., Tufto, J., Engen, S., Jerstad, K., Røstad, O. W. & Skåtan, J. E. Population dynamical consequences of climate change for a small temperate songbird. *Science* **287,** 854–856 (2000).
8. Sæther, B.-E. Environmental stochasticity and population dynamics of large herbivores: a search for mechanisms. *Trends Ecol. Evol.* **12,** 143–149 (1997).
9. Post, E., Stenseth, N. C., Langvatn, R. & Fromentin, J.-M. Global climate change and phenotypic variation among red deer cohorts. *Proc. R. Soc. Lond. B* **264,** 1317–1324 (1997).
10. Loison, A., Langvatn, R. & Solberg, E. J. Body mass and winter mortality in red deer calves: Disentangling sex and climate effects. *Ecography* **22,** 20–30 (1999).
11. Mysterud, A., Yoccoz, N. G., Stenseth, N. C. & Langvatn, R. Relationships between sex ratio, climate and density in red deer: the importance of spatial scale. *J. Anim. Ecol.* **69,** 959–974 (2000).
12. Lamb, P. J. & Peppler, R. A. North Atlantic Oscillation: concept and an application. *Am. Met. Soc.* **68,** 1218–1225 (1987).
13. Inouye, D. W., Barr, B., Armitage, K. B. & Inouye, B. D. Climate change is affecting altitudinal migrants and hibernating species. *Proc. Natl Acad. Sci. USA* **97,** 1630–1633 (2000).
14. Mysterud, A. Diet overlap among ruminants in Fennoscandia. *Oecologia* **124,** 130–137 (2000).
15. Hobbs, N. T. Linking energy balance to survival in mule deer: development and test of a simulation model. *Wildl. Monogr.* **101,** 1–39 (1989).
16. Albon, S. D. & Clutton-Brock, T. H. in *Ecological Change in the Uplands* (eds Usher, M. B. & Thompson, D. B. A.) 93–107 (Blackwell Scientific, Oxford, 1988).
17. Clutton-Brock, T. H. Guinness, F. E. & Albon, S. D. *Red Deer: Behaviour and Ecology of Two Sexes* (Edinburgh Univ. Press, Edinburgh, 1982).
18. Albon, S. D., Clutton-Brock, T. H. & Guinness, F. E. Early development and population dynamics in red deer. II. Density-independent effects and cohort variation. *J. Anim. Ecol.* **56,** 69–81 (1987).
19. Albon, S. D. & Langvatn, R. Plant phenology and the benefits of migration in a temperate ungulate. *Oikos* **65,** 502–513 (1992).
20. Van Soest, P. J. *Nutritional Ecology of the Ruminant* (Cornell Univ. Press, New York, 1994).
21. Kettlewell, P. S., Sothern, R. B. & Koukkari, W. L. U.K. wheat quality and economic value are dependent on the North Atlantic Oscillation. *J. Cereal Sci.* **29,** 205–209 (1999).
22. White, R. G. Foraging patterns and their multiplier effects on productivity of northern ungulates. *Oikos* **40,** 377–384 (1983).
23. Gaillard, J.-M., Festa-Bianchet, M., Delorme, D. & Jorgenson, J. Body mass and individual fitness in female ungulates: bigger is not always better. *Proc. R. Soc. Lond. B* **267,** 471–477 (2000).
24. Langvatn, R., Albon, S. D., Burkey, T. & Clutton-Brock, T. H. Climate, plant phenology and variation in age at first reproduction in a temperate herbivore. *J. Anim. Ecol.* **65,** 653–670 (1996).
25. Mysterud, A., Yoccoz, N. G., Stenseth, N. C. & Langvatn, R. The effects of age, sex and density on body weight of Norwegian red deer: evidence of density-dependent senescence. *Proc. R. Soc. Lond. B* **268** (in the press).
26. Langvatn, R. & Albon, S. D. Geographic clines in body weight of Norwegian red deer: a novel explanation of Bergmann's rule? *Holarct. Ecol.* **9,** 285–293 (1986).
27. Mysterud, A., Steinheim, G., Yoccoz, N. G., Holand, Ø. & Stenseth, N. C. Early onset of reproductive senescence in domestic sheep (*Ovis aries*): the cost of artificial selection? (submitted).
28. Hastie, T. & Tibshirani, R. *Generalized Additive Models* (Chapman and Hall, London, 1990).
29. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S-plus* (Springer, New York, 1994).
30. Post, E., Langvatn, R., Forchhammer, M. C. & Stenseth, N. C. Environmental variation shapes sexual dimorphism in red deer. *Proc. Natl Acad. Sci. USA* **96,** 4467–4471 (1999).

Acknowledgements

We gratefully acknowledge the financial support of the Research Council of Norway to A.M. and G.S.

Correspondence and request for material should be addressed to N.C.S. (e-mail: n.c.stenseth@bio.uio.no).

The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5

Fumitaka Hayashi*, Kelly D. Smith†‡, Adrian Ozinsky†, Thomas R. Hawn†§, Eugene C. Yi†, David R. Goodlett†, Jimmy K. Eng†, Shizuo Akira‖, David M. Underhill† & Alan Aderem†

* *Department of Immunology, ‡Department of Pathology, §Division of Allergy and Infectious Diseases, University of Washington, Seattle, Washington 98195, USA*

† *Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, Washington 98195, USA*

‖ *Department of Host Defense, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamada-oka, Suita, Osaka. 565-0871, Japan*

The innate immune system recognizes pathogen-associated molecular patterns (PAMPs) that are expressed on infectious agents, but not on the host. Toll-like receptors (TLRs) recognize PAMPs and mediate the production of cytokines necessary for the development of effective immunity[1–4]. Flagellin, a principal component of bacterial flagella, is a virulence factor that is recognized by the innate immune system in organisms as diverse as flies, plants and mammals[5–11]. Here we report that mammalian TLR5 recognizes bacterial flagellin from both Gram-positive and Gram-negative bacteria, and that activation of the receptor mobilizes the nuclear factor NF-κB and stimulates tumour necrosis factor-α production. TLR5-stimulating activity was purified from *Listeria monocytogenes* culture supernatants and identified as flagellin by tandem mass spectrometry. Expression of *L. monocytogenes* flagellin in non-flagellated *Escherichia coli* conferred on the bacterium the ability to activate TLR5, whereas deletion of the flagellin genes from *Salmonella typhimurium* abrogated TLR5-stimulating activity. All known TLRs signal through the adaptor protein MyD88. Mice challenged with bacterial flagellin rapidly produced systemic interleukin-6, whereas MyD88-null mice did not respond to flagellin. Our data suggest that TLR5, a member of the evolutionarily conserved Toll-like receptor family, has evolved to permit mammals specifically to detect flagellated bacterial pathogens.

© 2001 Macmillan Magazines Ltd

The Toll-like receptor family is defined by the presence of a conserved cytoplasmic signalling domain, the Toll/interleukin (IL)-1 receptor homology domain (TIR)[12]. TLR5 was identified by the presence of this domain[13–15], and is expressed in monocytes, immature dendritic cells, and epithelial cells[16,17]. TLR5's homology to other TLRs, and expression in immune cells, suggest that it may be a receptor involved in immune responses. Chimaeric molecules bearing the extracellular domain of CD4 fused to the transmembrane and TIR domains of TLR4 are constitutively active because of CD4-induced dimerization[18,19]. We found that a similar constitutively active form of TLR5 strongly activated NF-κB in Chinese hamster ovary (CHO) cells to an extent identical to that of the described CD4–TLR4 fusion (Fig. 1a)[18,19]. The proinflammatory nature of the TLR5 signal was confirmed by the finding that expression of CD4–TLR5 induced tumour necrosis factor-α (TNF-α) production in mouse macrophages (Fig. 1b). Thus, homo-oligomerization of the TLR5 signalling domain induces a proinflammatory signal that is related to that of TLR4.

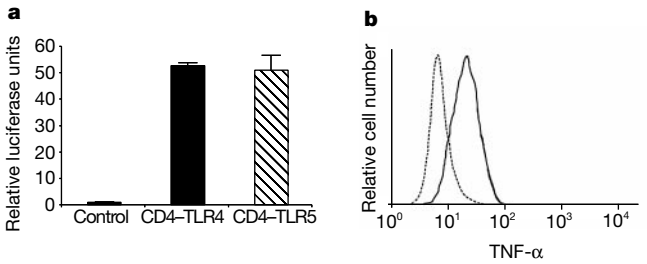


Figure 1 Constitutively active TLR5 activates NF-κB and TNF-α production. **a**, CHO cells transiently transfected with expression vectors for CD4–TLR4, CD4–TLR5, or empty expression vector (control), together with an NF-κB luciferase reporter. NF-κB-induced luciferase activity was measured. **b**, Transfection of RAW TT10 macrophages with a CD4–TLR5 expression vector. The production of TNF-α was measured by flow cytometry. The dotted line indicates TNF-α produced in cells not expressing CD4–TLR5; solid line indicates TNF-α produced in cells expressing CD4–TLR5.

CHO cells expressing human TLR5 and a luciferase-linked reporter were used to screen for PAMPs recognized by the receptor. TLR5 did not respond to any of the PAMPs known to stimulate TLR pathways, such as lipopolysaccharide (LPS), lipopeptide, yeast cell wall or peptidoglycan; whereas CHO cells transfected with TLR2 were stimulated by lipopeptide, yeast cell wall and peptidoglycan (Fig. 2a; and data not shown). However, TLR5-stimulating activity was detected in culture supernatants of a variety of Gram-positive and Gram-negative bacteria (Fig. 2b; and data not shown). The TLR5-stimulating activity of Gram-positive bacteria was not enhanced by co-expression of CD14 (data not shown). Notably, the TOP10 strain of *E. coli* had very little TLR5 activity (Fig. 2b), and was used in subsequent reconstitution experiments (see below). Experiments using murine TLR5 yielded similar results (data not shown).

The biological activity recognized by TLR5 was determined to be trichloroacetic-acid precipitable, phenol soluble, and sensitive to proteinase K and trypsin digestion (data not shown). To identify the bacterial components that stimulate TLR5, the supernatant from a saturated *L. monocytogenes* culture was concentrated and fractionated by reverse-phase chromatography, and each fraction was

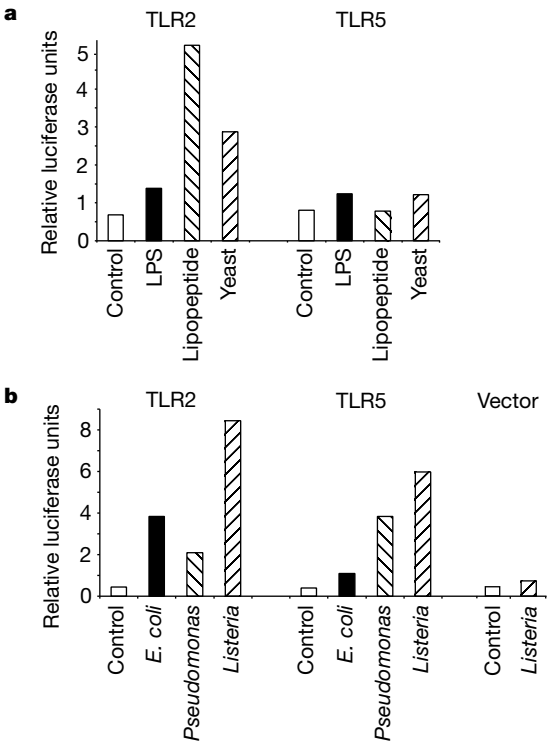


Figure 2 Bacterial culture supernatants contain TLR5-stimulating activity. CHO cells were transiently transfected with TLR2, TLR5, or empty expression vectors together with a NF-κB luciferase reporter. **a**, Luciferase activity in cells treated with 100 ng ml^{-1} LPS, 100 ng ml^{-1} lipopeptide, 10^7 yeast particles per ml, or untreated (control) cells. **b**, Luciferase activity in cells treated with 67 μl ml^{-1} of supernatant from the indicated saturated bacterial cultures or LB alone (control). Data are representative of three independent experiments.

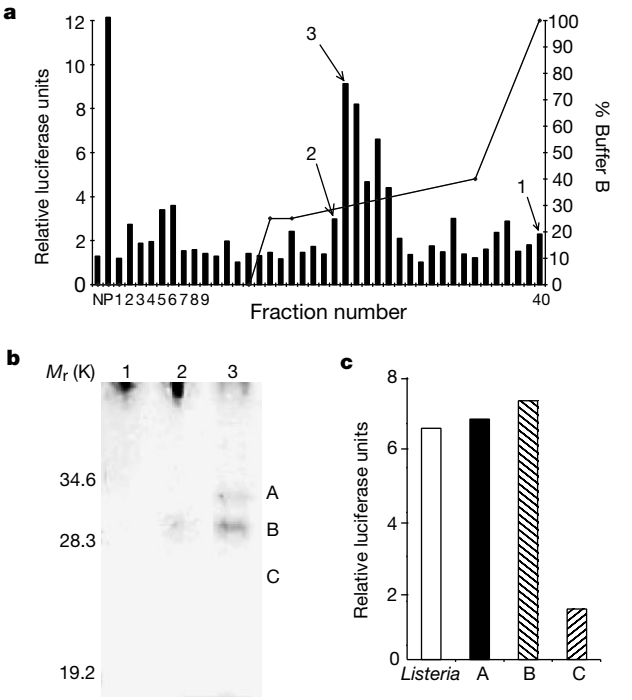


Figure 3 Purification of TLR5-stimulating activity from *L. monocytogenes* culture supernatant. **a**, NF-κB luciferase activity mediated by TLR5 in CHO cells expressing an NF-κB luciferase reporter and TLR5. Cells were stimulated with reverse-phase FPLC fractions. N, control LB growth medium; P, *L. monocytogenes* culture supernatant before chromatography. The fraction numbers correspond to 3-min fractions of reverse-phase FPLC eluted with the non-linear gradient of buffer B as shown. **b**, SDS–page and silver stain of fractions. Lane 1, fractions with no activity; 2, low activity; and 3, high activity, as indicated in **a**. **c**, Proteins eluted from regions A, B and C (indicated in **b**) assayed for NF-κB activation mediated by TLR5 in CHO cells. *Listeria* indicates *L. monocytogenes* culture supernatant. TLR5-stimulating activity was not recovered from any other section of the gel (data not shown).

© 2001 Macmillan Magazines Ltd

assessed for TLR5-stimulating activity in CHO cells (Fig. 3a). Fractions containing different amounts of TLR5-stimulating activity (Fig. 3a, arrows) were subjected to SDS–polyacrylamide gel electrophoresis (SDS–PAGE) (Fig. 3b). Two bands with apparent relative molecular masses of 30,000–34,000 (M_r 30K–34K) were clearly enriched in the fraction containing the highest level of TLR5-stimulating activity (Fig. 3b, lane 3). The bands were excised from the gel, eluted, and found to contain TLR5-stimulating activity that was not present in other regions of the gel lane (Fig. 3c). One of these bands, (Fig. 3b, band A), was trypsin-treated, subjected to micro-capillary high-performance-liquid-choromatography (HPLC) tandem mass spectrometry, and identified by comparison of peptide tandem mass spectra to sequences in a non-redundant protein database using the computer program SEQUEST[20] (Fig. 4a). Five sequences were identified (Fig. 4a) that correspond to flagellin, the product of the *flaA* gene of *L. monocytogenes*, and their location within the protein is indicated in Fig. 4b. Band B is also flagellin, which migrates as a doublet of roughly 30K on SDS–PAGE (Fig. 3b)[21]. The observation that flagellin is the TLR5 ligand is also supported by the finding that the flagellated bacteria, *L. monocytogenes* and *Pseudomonas aeruginosa*, stimulate TLR5, whereas the TOP10 strain of *E. coli*, which has lost its flagella, does not (Fig. 2b). Similarly, TLR5-stimulating activity was found in *Bacillus subtilis*, *Listeria innocua*, *S. typhimurium* and *Salmonella minnesota*, all flagellated bacteria, although non-flagellated bacteria such as *Haemophilus influenza* did not activate TLR5 (data not shown). Finally, purified flagellin stimulated TLR5-expressing CHO cells, but not TLR2-expressing CHO cells (Fig. 4c).

It is possible that bacterial supernatants contain other TLR5-stimulating factors in addition to flagellin. To confirm that flagellin is the only TLR5 ligand in bacteria, *E. coli* (TOP10) that secrete little TLR5 activity (Fig. 2b) were transformed with the complementary DNA of *L. monocytogenes* flagellin (*flaA*) under the control of an inducible promoter. Supernatants of *E. coli* that were induced to express *L. monocytogenes flaA* contained substantial TLR5-stimulating activity (Fig. 5a), whereas supernatants from *E. coli* in which *flaA* expression was repressed, or from *E. coli* expressing *flaA* in the reverse orientation, contained little TLR5 activity (Fig. 5a). *Listeria monocytogenes* flagellin is not recognized by TLR2, as supernatants from *E. coli* that express *flaA* did not show enhanced TLR2-dependent stimulation of CHO cells relative to supernatants from *E. coli* with repressed *flaA* expression (Fig. 5b). In addition to these experiments that demonstrate reconstitution of TLR5-stimulating activity by the expression of flagellin, we next used a bacterium from which flagellin had been deleted, and observed that TLR5-stimulating activity was abrogated. *Salmonella typhimurium* possesses two genes for flagellin (*fljB* and *fliC*)[22]. Culture supernatants of *fljB– fliC+ S. typhimurium* contained TLR5-stimulating

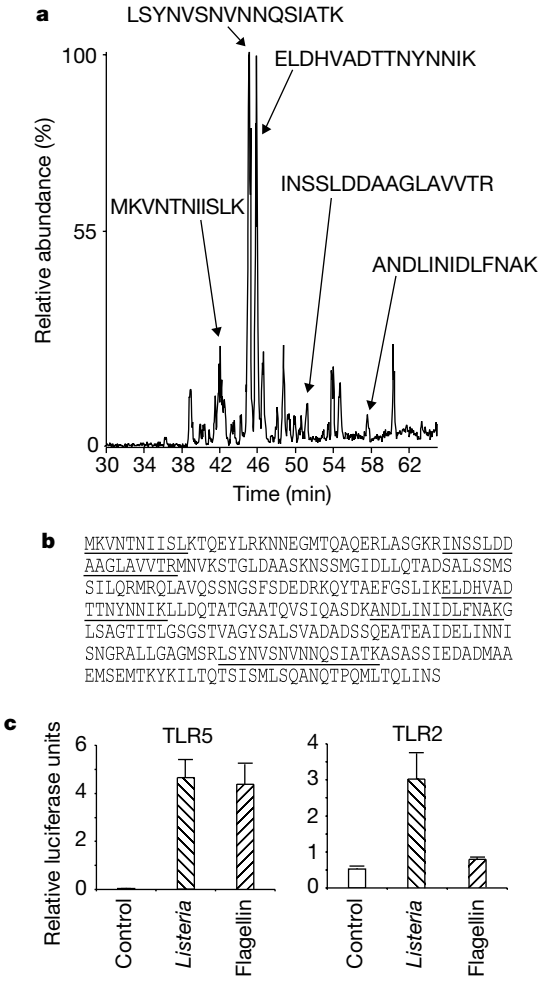


Figure 4 Identification by mass spectrometry of flagellin as the TLR5 stimulus. **a**, Microcapillary HPLC tandem mass spectrometry of Band A digested by trypsin. Peaks corresponding to *L. monocytogenes* flagellin peptides are indicated. **b**, Sequence of *L. monocytogenes* with indicated flagellin peptide positions (underlined) (GenBank number Q02551). **c**, Luciferase activity in CHO cells expressing an NF-κB luciferase reporter and TLR5 or TLR2, stimulated with 100 μl ml^{-1} *Listeria* supernatant or 33 μg ml^{-1} purified *Salmonella* flagellin. The mean and s.d. of quadruplicate samples are indicated.

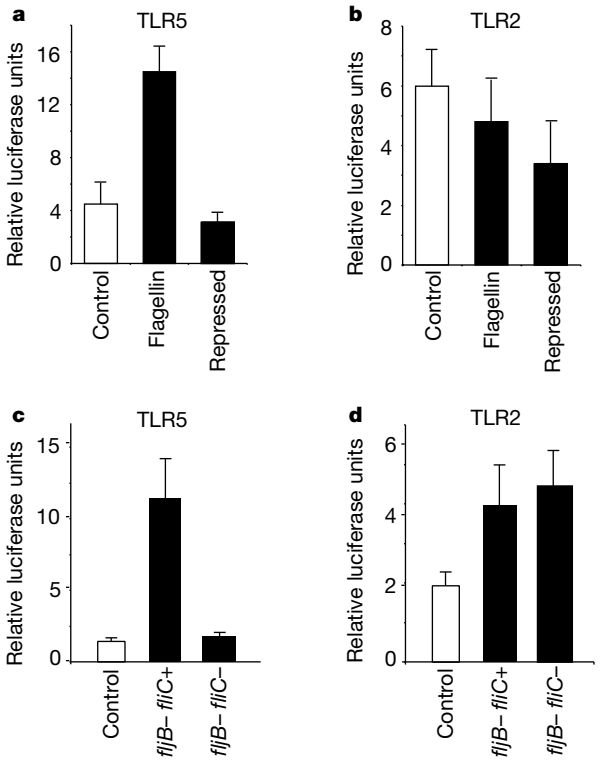


Figure 5 Flagellin expression in bacteria reconstitutes TLR5-stimulating activity, and deletion of flagellin genes abrogates TLR5-stimulating activity. **a**, **b**, CHO cells expressing an NF-κB luciferase reporter and TLR5 (**a**) or TLR2 (**b**) stimulated (5 h) with *E. coli* culture supernatants (67 μl ml^{-1}) in which expression of *L. monocytogenes* flagellin was induced or repressed. In the control sample, CHO cells were stimulated with supernatants from induced *E. coli* containing the *L. monocytogenes* flagellin gene cloned in the reverse orientation. **c**, **d**, CHO cells expressing an NF-κB luciferase reporter and TLR5 (**c**) or TLR2 (**d**), stimulated (5 h) with culture supernatants (100 μl ml^{-1}) from *S. typhimurium* lacking one copy of flagellin (*fljB– fliC+*) or both copies of flagellin (*fljB– fliC–*). Control, stimulation with LB medium. The mean and s.d. of quadruplicate samples are indicated.

© 2001 Macmillan Magazines Ltd

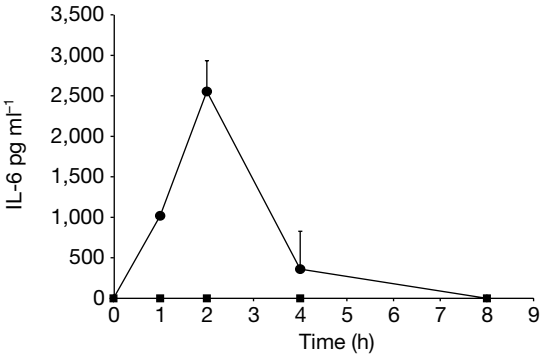


Figure 6 Flagellin-induced systemic IL-6 production is dependent on the TLR signalling adaptor molecule, MyD88. MyD88$^{-/-}$ mice (squares) or MyD88$^{+/+}$ littermate control mice (circles) were injected with 30 μg purified flagellin, and whole-blood IL-6 levels were monitored at the indicated time points. The data represents the mean for two mice sampled at 0, 1 and 8 h, and the mean and s.d. for three mice sampled at 2 and 4 h.

activity, whereas culture supernatants from *S. typhimurium* lacking both flagellins (*fljB*– *fliC*–) expressed no TLR5-stimulating activity (Fig. 5c). The lack of both flagellin genes had no effect on TLR2-stimulating activity (Fig. 5d). The observed TLR2-stimulating activity found in *S. typhimurium* supernatants is probably due to bacterial lipoproteins[23,24]. These results indicate that flagellin is the principal (and possibly the only) TLR5-stimulating activity present in *S. typhimurium* culture supernatant. Thus, TLR5-stimulating activity was elicited by introducing the flagellin gene into a non-flagellated bacterium, and abrogated by deleting the flagellin genes from a flagellated bacterium.

Next we examined whether TLR signalling is required for the *in vivo* immune response to flagellin. Intraperitoneal injection of wild-type mice with purified flagellin induced systemic IL-6 within 2 h (Fig. 6). By contrast, mice deficient in the adaptor protein MyD88, which is required for TLR signalling[1,3], were completely unresponsive to flagellin (Fig. 6).

Our study demonstrates that TLR5 is a pattern-recognition receptor and that its PAMP is flagellin, an observation consistent with the finding that the murine TLR5 gene lies within a locus that is associated with susceptibility to *Salmonella*. Flagellin stimulates host defence in a variety of organisms, including plants, insects and mammals[5–11]. This adjuvant property has been exploited by fusing polypeptides to flagellin to render them antigenic[25,26]. Clearly, the critical function of flagellin in bacterial motility places substantial constraints on the capacity of the molecule to mutate, and this renders it an ideal candidate for innate immune recognition. Indeed, flagellins have conserved amino and carboxy termini that are important in its function[27], and at least one of these domains is probably critical for TLR5 recognition—supported by our finding that TLR5 identifies flagellins from a variety of Gram-positive and Gram-negative bacteria. Thus, through recognition of flagellin TLR5 may serve as a general alarm system for a broad group of motile pathogens. □

Methods

Complementary DNA expression vectors

Human TLR5 and TLR2 were generated by PCR from cDNA derived from human peripheral blood mononuclear cells and cloned into pEF6-TOPO (Invitrogen) (pEF6::hTLR5 and pEF6::TLR2). Murine TLR5 was generated by PCR using cDNA derived from RAW TT10 cells[23] and cloned into pEF6 (pEF6::mTLR5). The construction of CD4–TLR4 has been described[19]. CD4–TLR5 was constructed by fusing the murine CD4 extracellular domain (residues 1–391) to the putative transmembrane and cytoplasmic domains of human TLR5 (residues 639–859) and cloned into pEF6-TOPO (pEF6::mCD4–hTLR5). *Listeria monocytogenes flaA* was generated by PCR from *L. monocytogenes* strain 10403, and cloned into pTrcHis2-TOPO (pTrcHis2::*flaA* (insert in sense orientation) and pTrcHis2::*flaA*rev (antisense orientation)) (Invitrogen).

CHO cell luciferase assays

CHO cells (CHO-K1) were obtained from ATCC (number CRL-9618) and grown in Hams F-12 medium supplemented with 10% FBS, L-glutamine, penicillin and streptomycin. For Figs 1 and 2, CHO cells were transfected by electroporation as described[23], with 1 μg of the indicated TLR expression vector, 1 μg of endothelial-leukocyte adhesion molecule (ELAM) firefly luciferase and 0.1 μg of thymidine kinase *Renilla* luciferase (Promega). Cells were plated on 96-well plates at 100,000 cells per well, and incubated overnight at 37 °C, 5% CO_2. The medium was replaced with media containing the stimuli at the indicated concentration/dilution. Bacterial lipopeptide, PAM_3CSK_4, was obtained from Roche, LPS (*S. minnesota* R595) was from List, and yeast particles (zymosan) were from Molecular Probes. Flagellin was purified from *S. typhimurium* (TH4778 (*fljB*–*fliC*+)) (see ref. 28 for procedure). The isolated flagellin was passed through a 100K MW-cutoff filter (Centricon-100, Millipore) to remove contaminating endotoxin. The purified flagellin contained 10 fg endotoxin per 1 ug flagellin protein, as determined by the limulus assay (BioWhittaker). Cells were stimulated for 5 h at 37 °C, and firefly and *Renilla* luciferase activities were measured using the Dual Luciferase Assay System (Promega). Luciferase activity is expressed as a ratio of NF-κB-dependent ELAM firefly luciferase activity divided by control thymidine kinase *Renilla* luciferase activity (relative luciferase units). For Figs 3–5, CHO cells were transfected as above with the addition of 0.1 μg pNeo/Tak[23], and stable populations of cells expressing the indicated TLR with the luciferase reporters were selected in 100 μg ml^{-1} G418. These cells were plated on 96-well plates at 100,000 cells per well, incubated overnight, and processed as above.

TNF-α assay

Flow cytometric measurements of TNF-α production by murine macrophages were conducted as described[19]. RAW TT10 cells were transfected with 10 μg pEF6::mCD4–hTLR5 by electroporation, and 18 h later the cells were incubated with 5 μg ml^{-1} of brefeldin A for 4 h to accumulate intracellular pools of newly synthesized TNF-α. Cells were fixed, permeabilized, stained for the expression of CD4 (anti-CD4-FITC, Pharmingen) and TNF-α (anti-murine TNF-α-phycoerythrin; Pharmingen), and analysed on a FACscan (Beckton–Dickenson). We analysed FACS data with WinMDI (J. Trotter, Scripps Research Institute, CA). Cells were gated to exclude dead cells and for expression of CD4.

Bacterial supernatants

Bacteria were grown either in Luria broth (LB): *E. coli* TOP10 (Invitrogen), *S. minnesota* (ATCC number 49284), mutant *S. typhimurium* (TH4778 (*fljB*– *fliC*+); TH2795 (*fljB*– *fliC*–); or in trypticase soy broth (TSB): *L. monocytogenes* (10403, a gift from D. Portnoy, UCSF), *L. innocua* (ATCC number 33090), *B. subtilis*, and *P. aeruginosa*. Bacteria were grown to saturation (about 16 h, 37 °C with vigorous aeration). The bacterial culture supernatants were centrifuged for 30 min at 2000*g*, filtered (0.2 μm), and stored at 4 °C before use. For *flaA* transfections, *E. coli* TOP10 containing pTrcHis2::*flaA* or pTrcHis2::*flaA*rev were selected from bacterial plates and grown to an optical density at 600 nm of 0.6 in LB with 100 μg ml^{-1} ampicillin and 1% w/v glucose. The bacteria were centrifuged for 30 min at 2000*g* and split into two LB cultures, one containing 100 μg ml^{-1} ampicillin and 1% w/v glucose (to repress *flaA*), and the other containing 100 μg ml^{-1} ampicillin and 1 mM IPTG (to induce *flaA*). Samples were taken at 4 h after induction, centrifuged for 5 min at 10,000*g*, and the supernatants stored at 4 °C before use.

Purification of the TLR5 ligand

A saturated *L. monocytogenes* culture (200 ml TSB) was centrifuged, and the supernatant was enriched for molecules larger than 30K by ultrafiltration (Ultrafree-15 filter unit with Biomax-30 membrane; Millipore). The buffer was changed to 100 mM Tris pH 7.5, and the volume was adjusted to 5 ml. The sample was loaded onto a HR5/10 reverse-phase chromatography column (AP Biotech) and run at 0.3 ml min^{-1}. Reverse-phase chromatography was performed with the indicated elution profile using the following buffers: (A) initial buffer, 0.1% TFA in water; (B) final buffer, 0.1% TFA in acetonitrile. FPLC fractions were separated on a 10% SDS–PAGE gel, and silver stained according to established protocols. TLR5-mediated activity was recovered from the gel as described below, omitting the trypsin digest.

Mass spectrometry

Proteins were separated by SDS–PAGE and visualized by silver staining. The band of interest was then excised, dehydrated with acetonitrile, dried under reduced vacuum, and trypsin (12.5 ng μl^{-1}) was infused into the gel. The gel slice was incubated on ice for 45 min in the presence of trypsin, the excess trypsin was removed and replaced with 50 mM ammonium bicarbonate, and the gel slice was incubated overnight at 37 °C. Peptides were extracted by three washes with 5% acetic acid in 50% aqueous acetonitrile. We pooled and concentrated the extractions by vacuum centrifugation. The peptides were injected onto a C18 peptide-trap cartridge (Michrom BioResources), desalted, and then injected onto a 75 μm (internal diameter) × 10 cm micro-capillary HPLC column (Magic C18; 5-μm packing, 100 Å pore size; Michrom BioResources). The sample injection was made using a FAMOS autosampler (LCPackings) coupled with an Agilent HP1100 Pump. Peptides were separated by a linear gradient of acetonitrile, and subjected to collision-induced dissociation using an electrospray ionization-ion-trap mass spectrometer (ESI-ITMS; ThermoQuest) in data-dependent mode with dynamic exclusion[29]. Protein identification was accomplished by use of the SEQUEST program[20].

Analysis of systemic immune responses to flagellin

MyD88$^{-/-}$mice (129/SvJ × C57Bl/6 background) were backcrossed for three generations with C57Bl/6 mice[30]. Mice from the F_3 generation (MyD88$^{-/-}$, $n = 5$) and littermate

© 2001 Macmillan Magazines Ltd

controls (MyD88$^{+/+}$, $n = 5$) were injected i.p. with 30 μg flagellin in 0.5 cc of saline. Blood was sampled at 0, 1, 2, 4 and 8 h after injection, and IL-6 levels were determined by ELISA (DuoSet, R&D Systems).

Received 8 December 2000; accepted 6 March 2001.

1. Aderem, A. & Ulevitch, R. J. Toll-like receptors in the induction of the innate immune response. *Nature* **406,** 782–787 (2000).
2. Brightbill, H. D. & Modlin, R. L. Toll-like receptors: molecular mechanisms of the mammalian immune response. *Immunology* **101,** 1–10 (2000).
3. Medzhitov, R. & Janeway, C. Jr Innate immune recognition: mechanisms and pathways. *Immunol. Rev.* **173,** 89–97 (2000).
4. Means, T. K., Golenbock, D. T. & Fenton, M. J. The biology of Toll-like receptors. *Cytokine Growth Factor Rev.* **11,** 219–232 (2000).
5. Samakovlis, C., Asling, B., Boman, H. G., Gateff, E. & Hultmark, D. *In vitro* induction of cecropin genes—an immune response in a *Drosophila* blood cell line. *Biochem. Biophys. Res. Commun.* **188,** 1169–1175 (1992).
6. Gomez-Gomez, L. & Boller, T. FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Mol. Cell* **5,** 1003–1011 (2000).
7. Wyant, T. L., Tanner, M. K. & Sztein, M. B. Potent immunoregulatory effects of *Salmonella typhi* flagella on antigenic stimulation of human peripheral blood mononuclear cells. *Infect. Immun.* **67,** 1338–1346 (1999).
8. Wyant, T. L., Tanner, M. K. & Sztein, M. B. *Salmonella typhi* flagella are potent inducers of proinflammatory cytokine secretion by human monocytes. *Infect. Immun.* **67,** 3619–3624 (1999).
9. Steiner, T. S., Nataro, J. P., Poteet-Smith, C. E., Smith, J. A. & Guerrant, R. L. Enteroaggregative *Escherichia coli* expresses a novel flagellin that causes IL-8 release from intestinal epithelial cells. *J. Clin. Invest.* **105,** 1769–1777 (2000).
10. McDermott, P. F., Ciacci-Woolwine, F., Snipes, J. A. & Mizel, S. B. High-affinity interaction between gram-negative flagellin and a cell surface polypeptide results in human monocyte activation. *Infect. Immun.* **68,** 5525–5529 (2000).
11. Ciacci-Woolwine, F., McDermott, P. F. & Mizel, S. B. Induction of cytokine synthesis by flagella from gram-negative bacteria may be dependent on the activation or differentiation state of human monocytes. *Infect. Immun.* **67,** 5176–5185 (1999).
12. Medzhitov, R. *et al.* MyD88 is an adaptor protein in the hToll/IL-1 receptor family signaling pathways. *Mol. Cell* **2,** 253–258 (1998).
13. Rock, F. L., Hardiman, G., Timans, J. C., Kastelein, R. A. & Bazan, J. F. A family of human receptors structurally related to *Drosophila* Toll. *Proc. Natl Acad. Sci. USA* **95,** 588–593 (1998).
14. Chaudhary, P. M. *et al.* Cloning and characterization of two Toll/Interleukin-1 receptor-like genes TIL3 and TIL4: evidence for a multi-gene receptor family in humans. *Blood* **91,** 4020–4027 (1998).
15. Sebastiani, G. *et al.* Cloning and characterization of the murine toll-like receptor 5 (Tlr5) gene: sequence and mRNA expression studies in Salmonella-susceptible MOLF/Ei mice. *Genomics* **64,** 230–240 (2000).
16. Muzio, M. *et al.* Differential expression and regulation of toll-like receptors (TLR) in human leukocytes: selective expression of TLR3 in dendritic cells. *J. Immunol.* **164,** 5998–6004 (2000).
17. Cario, E. & Podolsky, D. K. Differential alteration in intestinal epithelial cell expression of Toll-like receptor 3 (TLR3) and TLR4 in inflammatory bowel disease. *Infect. Immun.* **68,** 7010–7017 (2000).
18. Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A. Jr A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. *Nature* **388,** 394–397 (1997).
19. Ozinsky, A. *et al.* The repertoire for pattern recognition of pathogens by the innate immune system is defined by cooperation between Toll-like receptors. *Proc. Natl Acad. Sci. USA* **97,** 13766–13771 (2000).
20. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
21. Peel, M., Donachie, W. & Shaw, A. Temperature-dependent expression of flagella of *Listeria monocytogenes* studied by electron microscopy, SDS-PAGE and western blotting. *J. Gen. Microbiol.* **134,** 2171–2178 (1988).
22. Fujita, H., Yamaguchi, S. & Iino, T. Studies on H-O variants in Salmonella in relation to phase variation. *J. Gen. Microbiol.* **76,** 127–134 (1973).
23. Underhill, D. M. *et al.* The Toll-like receptor 2 is recruited to macrophage phagosomes and discriminates between pathogens. *Nature* **401,** 811–815 (1999).
24. Brightbill, H. D. *et al.* Host defense mechanisms triggered by microbial lipoproteins through toll-like receptors. *Science* **285,** 732–736 (1999).
25. das Gracas Luna, M., Sardella, F. F. & Ferreira, L. C. Salmonella flagellin fused with a linear epitope of colonization factor antigen I (CFA/I) can prime antibody responses against homologous and heterologous fimbriae of enterotoxigenic *Escherichia coli*. *Res. Microbiol.* **151,** 575–582 (2000).
26. Stocker, B. A. & Newton, S. M. Immune responses to epitopes inserted in Salmonella flagellin. *Int. Rev. Immunol.* **11,** 167–178 (1994).
27. Mimori-Kiyosue, Y., Vonderviszt, F. & Namba, K. Locations of terminal segments of flagellin in the filament structure and their roles in polymerization and polymorphism. *J. Mol. Biol.* **270,** 222–237 (1997).
28. Ibrahim, G. F., Fleet, G. H., Lyons, M. J. & Walker, R. A. Method for the isolation of highly purified Salmonella flagellins. *J. Clin. Microbiol.* **22,** 1040–1044 (1985).
29. Goodlett, D. R. *et al.* Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. *Anal. Chem.* **72,** 1112–1118 (2000).
30. Adachi, O. *et al.* Targeted disruption of the MyD88 gene results in loss of IL-1- and IL-18-mediated function. *Immunity* **9,** 143–150 (1998).

Acknowledgements

We thank D. Portnoy, S. R. Swanzy and K. T. Hughes for reagents and discussions; P. deRoos for advice with chromatography; and H. Harowicz for help with animal studies. F.H. and T.R.H. received support from the HHMI. K.D.S. received support from the NIH. This research was funded by NIH grants to A.A.

Correspondence and requests for materials should be addressed to A.A. (e-mail: aderem@systemsbiology.org).

TREM-1 amplifies inflammation and is a crucial mediator of septic shock

Axel Bouchon*, Fabio Facchetti†‡, Markus A. Weigand‡§ & Marco Colonna*

** Basel Institute for Immunology, Grenzacherstrasse 487, CH-4005 Basel, Switzerland*
† Institute of Pathology, University of Brescia, Spedali Civili 1, Brescia, Italy
§ Clinic of Anaesthesiology, University of Heidelberg and Tumor Immunology Program, German Cancer Research Institute, (DKFZ), INF 110,280, 69120 Heidelberg, Germany
‡ These authors contributed equally to this work

Host innate responses to bacterial infections are primarily mediated by neutrophils and monocytes/macrophages[1,2]. These cells express pattern recognition receptors (PRRs) that bind conserved molecular structures shared by groups of microorganisms[3,4]. Stimulation of PRR signalling pathways initiates secretion of proinflammatory mediators[3,4], which promote the elimination of infectious agents and the induction of tissue repair. Excessive inflammation owing to bacterial infections can lead to tissue damage and septic shock[5–9]. Here we show that inflammatory responses to microbial products are amplified by a pathway mediated by triggering receptor expressed on myeloid cells (TREM)-1. TREM-1 is an activating receptor expressed at high levels on neutrophils and monocytes that infiltrate human tissues infected with bacteria. Furthermore, it is upregulated on peritoneal neutrophils of patients with microbial sepsis and mice with experimental lipopolysaccaride (LPS)-induced shock. Notably, blockade of TREM-1 protects mice against LPS-induced shock, as well as microbial sepsis caused by live *Escherichia coli* or caecal ligation and puncture. These results demonstrate a critical function of TREM-1 in acute inflammatory responses to bacteria and implicate TREM-1 as a potential therapeutic target for septic shock.

We have identified recently a new receptor of the immunoglobulin superfamily expressed on human neutrophils and monocytes, called TREM-1 (ref. 10), which promotes cell activation through an associated signal transduction molecule DAP12 (ref. 11). The ability of TREM-1 to trigger secretion of proinflammatory mediators[10] prompted us to investigate its role in inflammation caused by bacteria. We determined TREM-1 expression by flow cytometry on neutrophils and monocytes incubated *in vitro* with heat-inactivated Gram-positive bacteria, Gram-negative bacteria, mycobacteria, or with bacterial cell-wall components. TREM-1 expression was strongly upregulated by extracellular bacteria, such as *Pseudomonas aeruginosa* and *Staphylococcus aureus*, as well as lipoteichoic acid (LTA) and LPS (Fig. 1a–b). In contrast, intracellular bacteria, such as *Bacillus* Calmette-Guerin (BCG), and mycolic acid had no effect. These results indicate that TREM-1 is upregulated in the presence of extracellular bacteria. On ligation, TREM-1 synergized with LPS to induce secretion of proinflammatory cytokines (tumour necrosis factor (TNF)-α and interleukin (IL)-1β) in monocytes (Fig. 1c). Thus, TREM-1 signalling amplifies responses to microbial products.

In vivo TREM-1 expression was determined in tissue specimens derived from acute or granulomatous inflammatory lesions caused by bacterial, fungal or non-microbial agents. TREM-1 was highly expressed in neutrophils associated with skin lesions caused by *S. aureus*, such as folliculitis and impetigo (Fig. 2a–d). In addition, increased TREM-1 expression was observed in neutrophils associated with granulomatous lymphadenitides caused by *Bartonella henselae* and *Aspergillus fumigatus* (Fig. 2e–h). In the latter, TREM-1 was also expressed in epithelioid and multinucleated giant cells

© 2001 Macmillan Magazines Ltd

surrounding the granulomas (Fig. 2g). In contrast, TREM-1 was hardly detectable in non-microbial inflammations, such as psoriasis, ulcerative colitis and vasculitis caused by immune complexes, despite a considerable infiltration of neutrophils and monocytes (Fig. 3). These results are consistent with a predominant role of TREM-1 in acute and granulomatous inflammations caused by extracellular bacteria and fungi.

Excessive inflammatory response to infectious agents can lead to septic shock[5–9], which is characterized by massive release of pro-inflammatory mediators, such as TNF-α, IL-1β, macrophage migration inhibitory factor (MIF) and high mobility group (HMG)-1 protein. This process leads not only to tissue damage, but also to haemodynamic changes, multiple organ failure, and ultimately death[6,12–15]. Consistent with the function of TREM-1 in response to bacterial infections, TREM-1 surface expression was strongly increased on infiltrating neutrophils isolated from the peritoneal cavity of patients with septic shock due to bacterial peritonitis (Fig. 4a, right panel). In contrast, peritoneal lavage cells of patients with a systemic inflammatory response syndrome (SIRS) caused by non-microbial peritoneal inflammation showed normal levels of TREM-1 (Fig. 4a, left panel).

To assess the direct involvement of TREM-1 in inflammatory responses, we tested whether a TREM-1 receptor decoy reduces inflammation and lethal shock in murine models of sepsis. We cloned the murine homologue of human TREM-1 and generated a murine TREM-1-specific monoclonal antibody (data not shown). Using this antibody, we observed that murine TREM-1 expression was upregulated in peritoneal neutrophils during experimental LPS-induced shock (Fig. 4b). A fusion protein containing murine TREM-1 extracellular domain and human immunoglobulin-γ (IgG1) Fc portion (mTREM-1/IgG1) was produced and injected in the peritoneal cavity 1 h before the induction of endotoxaemia. Lethality was monitored over time and compared with animals that had received control injections of human IgG1 (hIgG1), control IgG1 fusion protein (ILT-3/IgG1)[16] or heat-inactivated mTREM-1/IgG1 before LPS administration. As shown in Fig. 5a, 76% of the mice treated with mTREM-1/IgG1 survived endotoxaemia as compared with 6% of control mice. To quantify the protection provided by mTREM-1/IgG1, groups of mice pre-treated with mTREM-1/IgG1 or hIgG1 were challenged with various doses of LPS. The median lethal dose (LD_{50}) of LPS in animals treated with mTREM-1/IgG1 (LD_{50} = 621 μg) was significantly higher than the LD_{50} in control animals (LD_{50} = 467 μg) (Fig. 5b). We further monitored whether mTREM-1/IgG1 was still protective when administered 1,

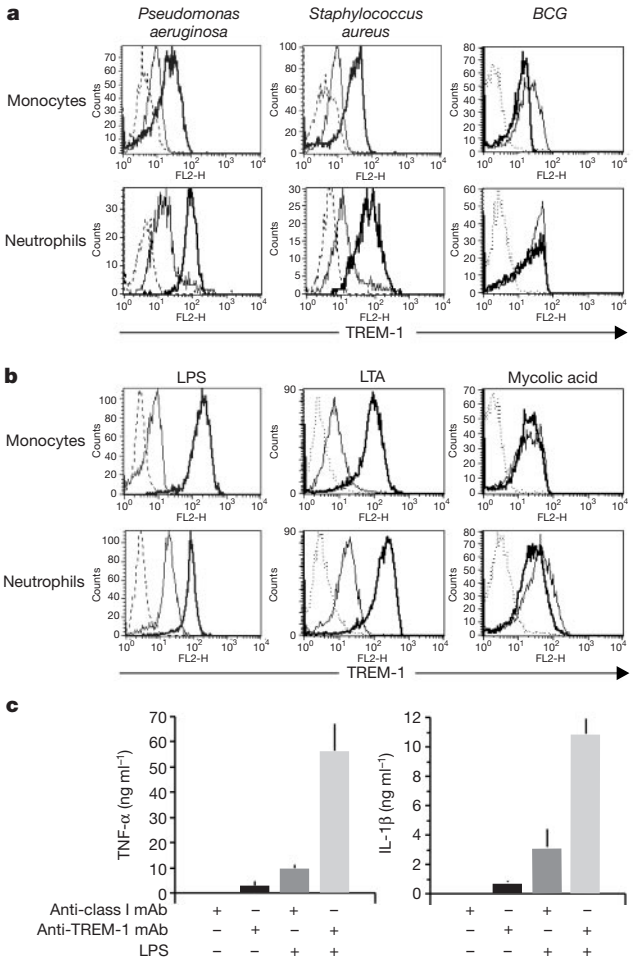


Figure 1 Regulation of human TREM-1 surface expression and function *in vitro*. **a**, **b**, TREM-1 is strongly upregulated after incubation of neutrophils and monocytes with heat-inactivated *S. aureus*, *P. aeruginosa*, LTA and LPS (left and middle panels) but not by BCG and mycolic acid (right panels). Stimulated (solid bold line) and non-stimulated (solid line) cells were analysed for cell-surface expression of TREM-1 (FL2-H). Dashed lines indicate background staining of stimulated cells with a control IgG1 monoclonal antibody (mAb). **c**, Ligation of TREM-1 potentiates LPS-mediated cytokine release. Monocytes were challenged for 16 h, as indicated. Supernatants were analysed for TNF-α (left panel) and IL-1β (right panel). All data points correspond to the mean and the s.d. of four independent experiments.

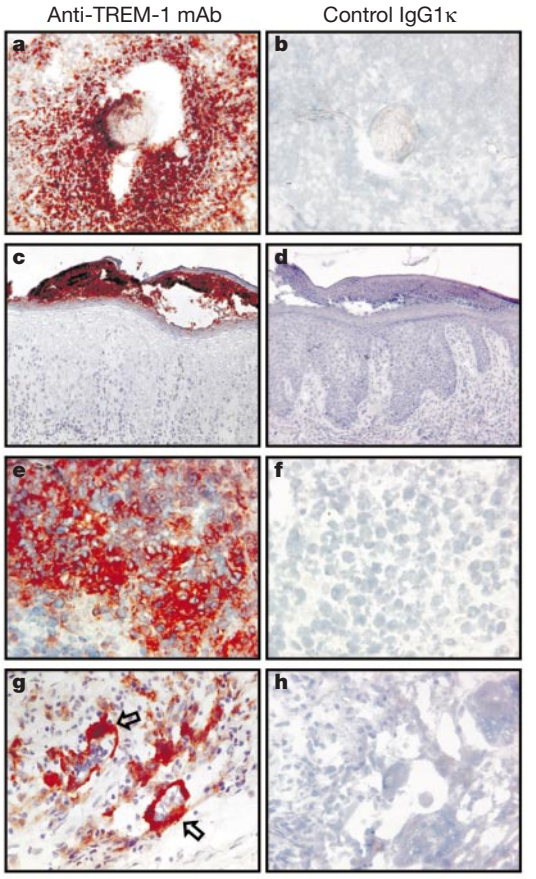


Figure 2 Human TREM-1 is strongly expressed in acute inflammatory lesions caused by bacteria and fungi. Staining with control mAb (**b**, **d**, **f**, **h**) and expression of TREM-1 (**a**, **c**, **e**, **g**). TREM-1 expression was detected using mAb 21C7 in acute cutaneous folliculitis (**a**) and impetigo (**c**) owing to *S. aureus*; in cat scratch granuloma induced by *B. henselae* (**e**); and in granuloma owing to *A. fumigatus* (**g**). Within the last (**g**), TREM-1 is expressed not only on infiltrating neutrophils but also on the multinucleated giant cells (arrows) surrounding the granuloma.

© 2001 Macmillan Magazines Ltd

2, 4 and 6 h after LPS injection. Notably, mTREM-1/IgG1 conferred 80% protection against endotoxic shock when applied 1 h after LPS injection. Partial protection was also observed after 2 and 4 h (Fig. 5c). Thus, soluble TREM-1 is effective even when injected after the outbreak of endotoxaemia.

Analysis of blood samples taken from mice pre-treated with mTREM-1/IgG1 and control animals at different time points after LPS administration showed a significant reduction of the plasma concentrations of both TNF-α and IL-1β (Fig. 5d, e). We also observed a significant reduction in the total cell number of neutrophils and monocytes/macrophages infiltrating the peritoneum 6–8 h after LPS injection in mTREM-1/IgG1-pre-treated animals as compared with controls (Fig. 5f, g). Injection of mTREM-1/IgG1 in normal mice did not affect the levels of circulating leukocytes. Furthermore, mTREM-1/IgG1 was effective against endotoxaemia even when the IgG1–Fc portion of the fusion protein was mutated to inhibit Fc receptor binding and complement fixation (data not shown). Thus, inhibition of TREM-1-mediated responses is sufficient to lower systemic levels of TNF-α and IL-1β, and to reduce cellular infiltrates at the site of inflammation below levels that are lethal for the host, without causing leukopenia.

Experimental endotoxic shock reproduces human sepsis only in part, as it does not involve the replication and dissemination of bacteria. In these conditions a complete block of TREM-1 signalling could be deleterious by impairing the capacity of the immune system to fight infections, as observed for anti-TNF-α treatments[17–23]. We therefore investigated whether mTREM-1/IgG1 protects against septic shock in two models of microbial peritonitis and sepsis caused by intraperitoneal (i.p.) administration of *E. coli* or by caecal ligation and puncture (CLP)[17,24,25]. As shown in Fig. 6, injection of mTREM-1/IgG1 conferred significant protection against lethal *E. coli* peritonitis and CLP-induced septic shock compared with control hIgG1, whereas treatment with TNF-α receptor I IgG1 (TNF-RI/IgG1) caused accelerated death of all animals (Fig. 6b). Thus, mTREM-1/IgG1 reduces inflammatory responses but allows sufficient control of the bacterial infection.

Our results demonstrate that TREM-1 is an amplifier of inflammatory responses that are triggered by bacterial and fungal infections. In the early phase of infection, neutrophils and monocytes initiate the inflammatory response owing to the engagement of PRRs by microbial products. At the same time, bacterial products induce upregulation of TREM-1. On recognition of an unknown ligand, TREM-1 activates DAP12-signalling pathways[10,11], which amplify inflammatory responses. We have shown that ligation of TREM-1 leads to sustained secretion of proinflammatory cytokines (TNF-α and IL-1β) and chemokines (IL-8 and monocyte chemotactic protein (MCP)-1)[10] and may also result in prolonged survival of neutrophils and monocytes at the inflammatory site. Inhibition of functions meditated by TREM-1 in LPS-induced shock is sufficient to reduce serum TNF-α and IL-1β to sublethal levels, preventing shock and death. Furthermore, mTREM-1/IgG1 protects against bacterial peritonitis, in contrast to treatment with TNF-RI/IgG1 (Fig. 6b) and anti-TNF-α antibodies[17], which increase lethality. The residual levels of TNF-α and IL-1β after mTREM-1/IgG1 treatment are probably sufficient for clearance of bacterial infections[17–23]. Murine TREM-1-IgG1 was even protective after injection of LPS, an effect that was only reported for inhibition of MIF and HMG-1[12,24]. Thus, administration of soluble TREM-1 after

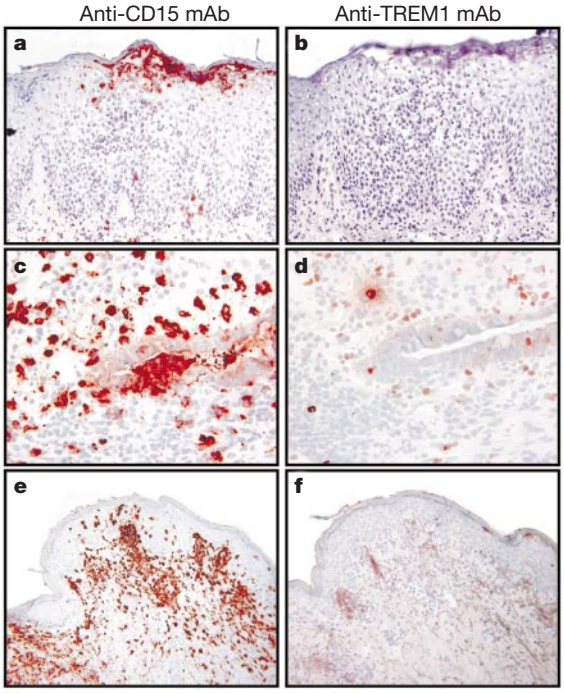


Figure 3 Human TREM-1 is only weakly expressed in neutrophils and monocytes accumulating in non-microbial inflammations. Psoriasis (**a**, **b**), ulcerative colitis (**c**, **d**), and vasculitis caused by immune complexes (**e**, **f**) are characterized by inflammatory infiltrates of neutrophils, as detected by anti-CD15 mAb (left panels). TREM-1 expression is weak or absent, however, in consecutive/serial sections (right panels).

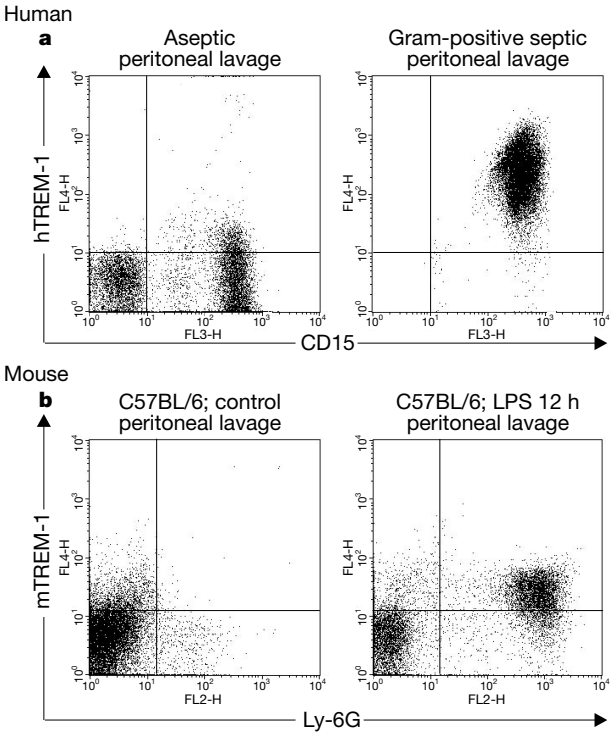


Figure 4 TREM-1 is strongly upregulated on peritoneal neutrophils during septic shock in humans and mice. **a**, Flow cytometric analysis of peritoneal lavage cells from patients with aseptic SIRS owing to aseptic cholecystitis (left) or polymicrobial Gram-positive sepsis caused by bowel perforation (right). $CD15^+$ cells correspond to neutrophils. **b**, Four-colour analysis of peritoneal leucocytes from LPS-treated C57BL/6 mice (right) compared with control animals (left) $Ly\text{-}6G^+/TREM\text{-}1^+$ cells correspond to murine neutrophils. The $Ly\text{-}6G^+/TREM\text{-}1^+$ cells are $CD11b/Mac\text{-}1^+$(data not shown) and therefore correspond to peritoneal macrophages. Staining with isotype-matched control monoclonal antibodies were set to the indicated lower quadrants.

© 2001 Macmillan Magazines Ltd

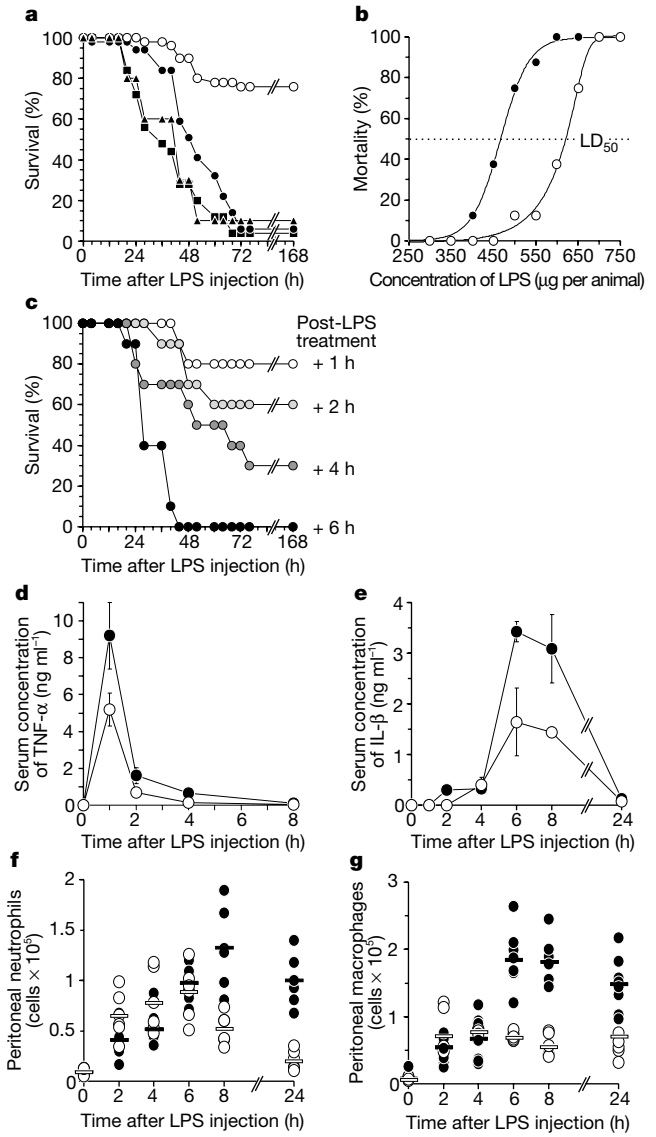


Figure 5 Inhibition of mTREM-1 signalling blocks endotoxic shock and inflammatory responses *in vivo*. **a**, C57BL/6 mice were treated with control hIgG1 (filled circles) or mTREM-1/IgG1 (open circles) 1 h before LPS administration. Data points are from seven independent experiments (5–10 animals per group). Survival was 76% (37 out of 49) in mice treated with mTREM-1/IgG1 and 6% (3 out of 49) in mice treated with hIgG1 (P = 0.0002, two-tailed Fisher's exact test). In additional controls, mice received injections with purified human ILT3-IgG1 (filled squares; n = 25) or heat-inactivated mTREM-1/IgG1 (filled triangles; n = 10) before induction of endotoxaemia. **b**, Estimation of LPS LD_{50} in mice treated with mTREM-1/IgG1 or hIgG1. Mice were randomly assigned to 20 groups, each containing 10 animals. Ten groups received i.p. injections of mTREM-1/IgG1, and 10 groups were injected with hIgG1. One hour later, endotoxaemia was induced by application of LPS, as indicated. Calculation of LD_{50} was accomplished as described (LD_{50} for mTREM-1/IgG1 = 621 μg; LD_{50} for hIgG1 = 467 μg; $P < 0.0001$)[6]. **c**, mTREM-1/IgG1 protects against LPS-induced lethal peritonitis when given after challenge. Mice were injected with LPS before administration of mTREM-1/IgG1. Data points are from two independent experiments (3–7 animals per group). Survival was 80%, 60% (P = 0.0007 and P = 0.0108, respectively; two-tailed Fisher's exact test), 40% and 0%, respectively. **d**–**g**, Analysis of inflammatory parameters during fatal endotoxaemia. Mice were treated as described in **a**. Serum levels of TNF-α (**d**) and IL-1β (**e**), and numbers of peritoneal neutrophils (**f**) and macrophages (**g**) were determined at the indicated time points. Data points correspond to the mean and the s.d. of two independent experiments (4–6 mice per treatment group).

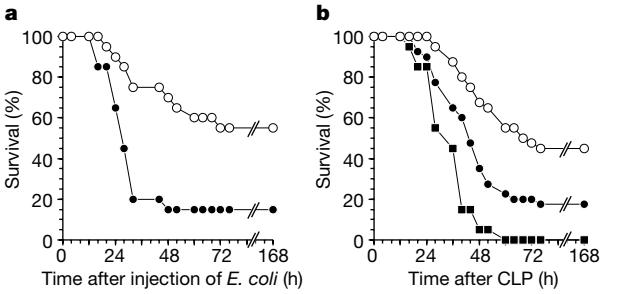


Figure 6 mTREM-1 is protective in bacterial peritonitis. **a**, C57BL/6 mice injected i.p. with mTREM-1/IgG1 (open circles) or hIgG1 (filled circles) 1 h before i.p. administration of *E. coli*. Data points are from two independent experiments (5–15 animals per group). Survival was 55% (11 out of 20) in mice treated with mTREM-1/IgG1 and 15% (3 out of 20) in mice treated with control hIgG1 (P = 0.0187, two-tailed Fisher's exact test). **b**, mTREM-1/IgG1 protects against fatal septic shock induced by CLP. Mice were injected i.p. with mTREM-1/IgG1 (open circles), hIgG1 (filled circles) or TNF-RI/IgG1 (filled squares) immediately after CLP. Data points are from four independent experiments (5–10 animals per group). Survival was 45% (18 out of 40) in mice treated with mTREM-1/IgG1, 17.5% (7 out of 40) in mice treated with control hIgG1 (P = 0.015, two-tailed Fisher's exact test) and 0% (0 out of 20) in mice treated with TNF-RI/IgG1.

infection might be a suitable therapeutic tool for the treatment of septic shock as well as other microbial-mediated diseases. □

Methods

Human cells and bacteria

Human peripheral blood mononuclear cells, neutrophils and $CD14^+$ monocytes were purified from peripheral blood of healthy donors as described[11]. Human peritoneal leukocytes were obtained from peritoneal lavage of patients diagnosed with aseptic systemic inflammatory response syndrome or polymicrobial sepsis, as defined in ref. 26. Cultures of *S. aureus*, *P. aeruginosa* and BCG were stopped in the logarithmic growth phase, washed twice in PBS and inactivated by incubating for 30 min at 80 °C.

Flow cytometric and functional analysis of human cells

Purified monocytes and neutrophils were cultured in the absence or presence of inactivated bacteria (monocytes/neutrophils: bacteria ratios were about 1/10–1/100), LPS (100 ng ml^{-1}), LTA (100 ng ml^{-1}) and mycolic acid (10 μg ml^{-1}). All cells were incubated with PBS or 20% human serum for 1 h on ice to block Fc receptors. After staining with either monoclonal antibody 21C7 (mouse IgG1, anti-TREM-1) or 1B7.11 (control mouse IgG1, anti-2,4,6 TNP; American Type Culture Collection), followed by human-adsorbed phycoerythrin (PE)-conjugated goat anti-mouse IgG (Southern Biotechnology Associates), cells were analysed on a flow cytometer. Four-colour analysis of human peritoneal leukocytes was performed using anti-TREM-1, CD15 (Immunotech), CD14 (Immunotech) and CD16 (Immunotech) monoclonal antibodies conjugated with allophycocyanin (APC), CyChrome, PE and fluorescein isothiocyaniate (FITC), respectively.

Immunohistochemistry

We took skin biopsies from *S. aureus* infections (with features of impetigo and folliculitis), psoriasis or leukocytoclastic vasculitis mediated by immunocomplexes. Colon biopsies were from patients with active ulcerative colitis. The infectious granulomas derived from lymph nodes or mediastinum were caused by *Mycobacterium tuberculosis*, *B. henselae* (cat scratch disease) or *A. fumigatus*; the latter derived from a patient affected by chronic granulomatous disease. We also analysed cases of sarcoidosis (lymph node) and a foreign-body giant-cell reaction associated with a vascular plastic prosthesis removed because of thrombosis. Staining with 21C7, anti-CD15 (Dako), or an isotype-matched (IgG1) antibody was performed on frozen sections. The primary antibody was detected by the streptavidine–biotin immunoperoxidase technique[27].

mTREM-1-fusion protein and antibodies

Complementary DNA of mTREM-1 (GenBank accession number NM 021406) was cloned by searching the GenBank expressed sequence database with the amino-acid sequence of human TREM-1. We produced mTREM-1/IgG1 as described[10]. Anti-mTREM-1-specific monoclonal antibody was produced by immunizing Lewis rats with soluble mTREM-1/IgG1 as described[10].

LPS-induced endotoxaemia

Female C57BL/6 mice (8–10 weeks, 19–22 g) were randomly grouped (5–10 mice per group) and injected i.p. with different concentrations of LPS from *E. coli* 055:B5 (Sigma), in a blinded fashion. We administered 500 μg per mouse of purified hIgG1 (κ-light chain)

© 2001 Macmillan Magazines Ltd

(Sigma), mTREM-1/IgG1, hILT-3/IgG1[16], or heat-inactivated mTREM-1/IgG1 (30 min, 95 °C) by i.p. injection at 1, 2, 4 and 6 h after or 1 h before LPS administration. We monitored viability of treated mice 4–6 times a day for at least 10 days.

Analysis of blood and peritoneal lavage

Blood (250 μl) was collected and serum TNF-α and IL-1β were determined using cytokine-specific enzyme-linked immunosorbent assays as per the manufacturer's protocol (R&D Systems). Total cell numbers of peritoneal lavage cells collected from treated and control mice were determined on a coulter counter. We performed differential counts according to standard morphological criteria on cytospin preparations stained with Giemsa/May-Grünwald solution (Sigma). Four-colour analysis of peritoneal leukocytes was performed after blocking Fc receptors (FcR blocking agent; Pharmingen) for 30 min, using anti-mTREM-1, anti-Ly-6G (Pharmingen), anti-Mac-1 (Pharmingen) monoclonal antibodies conjugated with APC, PE and FITC, respectively. Dead cells were excluded by staining with propidium iodide.

Escherichia coli peritonitis model

E. coli peritonitis was induced in mice as described[25]. Briefly, female C57BL/6 mice (8–10 weeks, 19–22 g) were weighed and randomly distributed into groups of 5–15 animals of equal body weight. Mice were injected i.p. with 500 μg of mTREM-1/IgG1 or control hIgG1 before i.p. administration of 500 μl of a suspension of *E. coli* O111:B4 ($1.6–2.1 \times 10^6$ colony forming units per mouse).

Caecal ligation and puncture

We performed CLP as described[17,24]. Briefly, female C57BL/6 mice (8–10 weeks, 19–22 g) were anaesthetized by i.p. administration of 75 mg per kg Ketanest (Parke Davies) and 16 mg per kg Rompun (Bayer AG) in 0.2 ml sterile pyrogen-free saline (B. Braun Melsungen AG). The caecum was exposed through a 1.0–1.5-cm abdominal midline incision and subjected to a 50–80% ligation of the distal half followed by a single puncture with a G23 needle. A small amount of stool was expelled from the punctures to ensure patency. The caecum was replaced into the peritoneal cavity and the abdominal incision closed in layers with 5/0 Prolene thread (Ethicon). Sterile saline (500 μl) containing 500 μg mTREM-1/IgG1, 500 μg hIgG1κ (Sigma) or 100 μg TNF-RI/IgG1 (Pharmingen) (together with 400 μg hIgG1κ; Sigma) was administered by i.p. injection immediately after CLP. The CLP was performed blinded to the identity of the treatment group. We assessed survival after CLP 4–6 times a day for at least 7 days.

Received 22 January; accepted 26 February 2001.

1. Medzhitov, R. & Janeway, C. Jr Innate immunity. *N. Engl. J. Med.* **343,** 338–344 (2000).
2. Hoffmann, J. A., Kafatos, F. C., Janeway, C. A. & Ezekowitz, R. A. Phylogenetic perspectives in innate immunity. *Science* **284,** 1313–1318 (1999).
3. Aderem, A. & Ulevitch, R. J. Toll-like receptors in the induction of the innate immune response. *Nature* **406,** 782–787 (2000).
4. Beutler, B. Endotoxin, toll-like receptor 4, and the afferent limb of innate immunity. *Curr. Opin. Microbiol.* **3,** 23–28 (2000).
5. Bone, R. C. The pathogenesis of sepsis. *Ann. Intern. Med.* **115,** 457–469 (1991).
6. Beutler, B., Milsark, I. W. & Cerami, A. C. Passive immunization against cachectin/tumor necrosis factor protects mice from lethal effect of endotoxin. *Science* **229,** 869–871 (1985).
7. Morrison, D. C. & Ryan, J. L. Endotoxins and disease mechanisms. *Annu. Rev. Med.* **38,** 417–432 (1987).
8. Tracey, K. J. *et al.* Shock and tissue injury induced by recombinant human cachectin. *Science* **234,** 470–474 (1986).
9. Glauser, M. P., Zanetti, G., Baumgartner, J. D. & Cohen, J. Septic shock: pathogenesis. *Lancet* **338,** 732–736 (1991).
10. Bouchon, A., Dietrich, J. & Colonna, M. Cutting edge: inflammatory responses can be triggered by TREM-1, a novel receptor expressed on neutrophils and monocytes. *J. Immunol.* **164,** 4991–4995 (2000).
11. Lanier, L. L., Corliss, B. C., Wu, J., Leong, C. & Phillips, J. H. Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. *Nature* **391,** 703–707 (1998).
12. Wang, H. *et al.* HMG-1 as a late mediator of endotoxin lethality in mice. *Science* **285,** 248–251 (1999).
13. Bernhagen, J. *et al.* MIF is a pituitary-derived cytokine that potentiates lethal endotoxaemia. *Nature* **365,** 756–759 (1993).
14. Ohlsson, K., Bjork, P., Bergenfeldt, M., Hageman, R. & Thompson, R. C. Interleukin-1 receptor antagonist reduces mortality from endotoxin shock. *Nature* **348,** 550–552 (1990).
15. Alexander, H. R., Doherty, G. M., Buresh, C. M., Venzon, D. J. & Norton, J. A. A recombinant human receptor antagonist to interleukin 1 improves survival after lethal endotoxemia in mice. *J. Exp. Med.* **173,** 1029–1032 (1991).
16. Cella, M. *et al.* A novel inhibitory receptor (ILT3) expressed on monocytes, macrophages, and dendritic cells involved in antigen processing. *J. Exp. Med.* **185,** 1743–1751 (1997).
17. Echtenacher, B., Falk, W., Mannel, D. N. & Krammer, P. H. Requirement of endogenous tumor necrosis factor/cachectin for recovery from experimental peritonitis. *J. Immunol.* **145,** 3762–3766 (1990).
18. Echtenacher, B., Mannel, D. N. & Hultner, L. Critical protective role of mast cells in a model of acute septic peritonitis. *Nature* **381,** 75–77 (1996).
19. Malaviya, R., Ikeda, T., Ross, E. & Abraham, S. N. Mast cell modulation of neutrophil influx and bacterial clearance at sites of infection through TNF-α. *Nature* **381,** 77–80 (1996).
20. Rothe, J. *et al.* Mice lacking the tumour necrosis factor receptor 1 are resistant to TNF- mediated toxicity but highly susceptible to infection by *Listeria monocytogenes*. *Nature* **364,** 798–802 (1993).
21. Pfeffer, K. *et al.* Mice deficient for the 55 kd tumor necrosis factor receptor are resistant to endotoxic shock, yet succumb to *L. monocytogenes* infection. *Cell* **73,** 457–467 (1993).
22. Peschon, J. J. *et al.* TNF receptor-deficient mice reveal divergent roles for p55 and p75 in several models of inflammation. *J. Immunol.* **160,** 943–952 (1998).
23. Eskandari, M. K. *et al.* Anti-tumor necrosis factor antibody therapy fails to prevent lethality after cecal ligation and puncture or endotoxemia. *J. Immunol.* **148,** 2724–2730 (1992).
24. Calandra, T. *et al.* Protection from septic shock by neutralization of macrophage migration inhibitory factor. *Nature Med.* **6,** 164–170 (2000).
25. Appelmelk, B. J. *et al.* Use of mucin and hemoglobin in experimental murine gram-negative bacteremia enhances the immunoprotective action of antibodies reactive with the lipopolysaccharide core region. *Antonie Van Leeuwenhoek* **52,** 537–542 (1986).
26. American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Crit. Care. Med.* **20,** 864–874 (1992).
27. Facchetti, F. *et al.* Suppurative granulomatous lymphadenitis. Immunohistochemical evidence for a B-cell-associated granuloma. *Am. J. Surg. Pathol.* **16,** 955–961 (1992).

Acknowledgements

We thank R. Breitkreutz and A. Benner for statistical analysis; O. Alebardi and S. Festa for technical assistance in immunohistochemistry; N. Schmitz for advice during mouse experiments; M. Cella, R. Ettinger, F. McBlane, M. Kopf, F. Sinigaglia and R. Torres for reviewing the manuscript; and P. Krammer for discussion. This work was partly supported by a BIOMED-2 grant to F.F. The Basel Institute for Immunology was founded and is supported by F. Hoffmann-La Roche, CH-4002 Basel.

Correspondence and requests for materials should be addressed to M.C. (e-mail: colonna@bii.ch).

IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity

Vijay Shankaran*, Hiroaki Ikeda*, Allen T. Bruce*, J. Michael White*, Paul E. Swanson*, Lloyd J. Old† & Robert D. Schreiber*

* *Department of Pathology and Immunology, Center for Immunology, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA*

† *Ludwig Institute for Cancer Research, New York Branch at Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA*

Lymphocytes were originally thought to form the basis of a 'cancer immunosurveillance' process that protects immunocompetent hosts against primary tumour development[1,2], but this idea was largely abandoned when no differences in primary tumour development were found between athymic nude mice and syngeneic wild-type mice[3–5]. However, subsequent observations that nude mice do not completely lack functional T cells[6,7] and that two components of the immune system—IFNγ[8,9] and perforin[10–12]—help to prevent tumour formation in mice have led to renewed interest in a tumour-suppressor role for the immune response. Here we show that lymphocytes and IFNγ collaborate to protect against development of carcinogen-induced sarcomas and spontaneous epithelial carcinomas and also to select for tumour cells with reduced immunogenicity. The immune response thus functions as an effective extrinsic tumour-suppressor system. However, this process also leads to the immunoselection of tumour cells that are more capable of surviving in an immunocompetent host, which explains the apparent paradox of tumour formation in immunologically intact individuals.

Age-matched female wild-type mice and immunodeficient mice with a targeted disruption of the recombination-activating gene-2 (RAG2) that is expressed only in lymphocytes[13], both on a pure 129/SvEv genetic background, were injected subcutaneously with 100 μg of the chemical carcinogen methylcholanthrene (MCA) and monitored for tumour development. $RAG2^{-/-}$ mice developed tumours earlier than wild-type mice and with greater frequency ($P < 0.01$). After 160 days, 9/15 $RAG2^{-/-}$ mice but only 2/15 wild-type mice formed MCA-induced tumours (Fig. 1a). Similar results

© 2001 Macmillan Magazines Ltd

were obtained in two additional experiments. In total, 30/52 RAG2$^{-/-}$ mice formed MCA-induced tumours compared to only 11/57 wild-type mice (Fig. 1b) ($P < 0.0001$). The increased tumour formation observed in RAG2$^{-/-}$ mice was comparable to that observed in 129/SvEv strain, IFNγ-insensitive mice that lack either the IFNγ receptor (IFNGR1) or STAT1, the transcription factor that plays a dedicated role in mediating signalling by the IFNγ and IFNα/β receptors[14,15] (12/20 and 17/30, respectively, versus 11/57 wild-type mice; $P < 0.001$); this data is consistent with our previous findings[9] (Fig. 1b). To assess the extent of collaboration between the lymphocyte- and IFNγ/STAT1-dependent tumour-suppressor mechanisms, we generated RAG2$^{-/-}$ × STAT1$^{-/-}$ mice (RkSk mice) and assessed their MCA sensitivity. Mice lacking both genes also showed increased susceptibility to MCA-induced carcinogenesis with 13/18 mice forming tumours compared to 11/57 wild-type mice ($P < 0.0001$), but did not display a significantly greater tumour incidence compared to mice that lacked either the RAG2 or STAT1 genes individually (Fig. 1b) ($P = 0.127$ and 0.140, respectively). All four groups of mice formed histologically indistinguishable sarcomas at the injection site. Similar results were obtained using male mice, which are known to be more susceptible to MCA-induced tumorigenesis. Thus, T, NKT and/or B cells are essential to suppress development of chemically induced tumours.

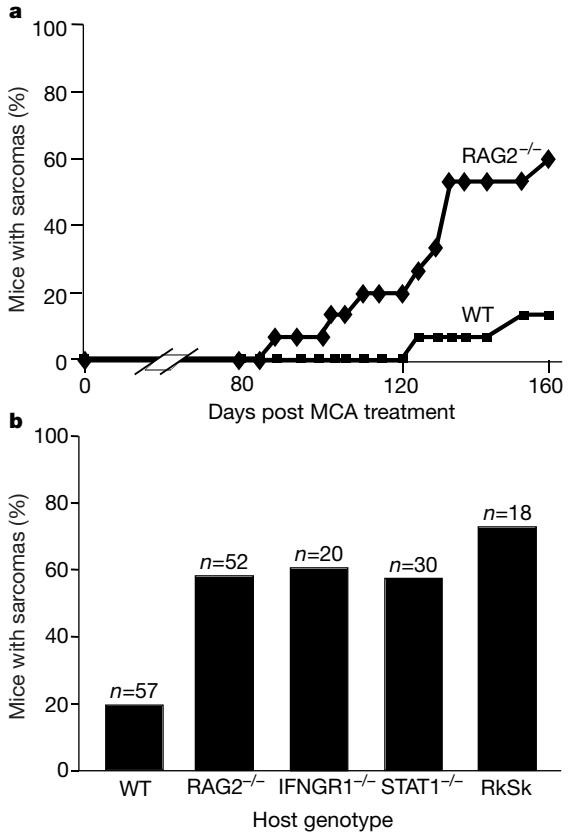


Figure 1 Lymphocyte-deficient mice are highly susceptible to MCA-induced tumour development. **a**, Fifteen female RAG2$^{-/-}$ mice (diamonds) and 15 syngeneic 129/SvEv wild-type (WT) mice (squares) were injected with a single subcutaneous dose of 100 μg MCA and observed for tumour development for 160 days. This dose and observation period were determined to be the most informative based on previous carcinogenesis experiments performed on the 129/SvEv strain of mice[9]. Data are represented as a percentage of each group of mice bearing tumours at least 9 mm in diameter as a function of time. **b**, Cumulative tumour formation at 160 days from three independent experiments involving female wild-type 129/SvEv mice ($n = 57$), RAG2$^{-/-}$ mice ($n = 52$), STAT1$^{-/-}$ mice ($n = 30$), IFNGR1$^{-/-}$ mice ($n = 20$) and RAG2$^{-/-}$ × STAT1$^{-/-}$ (RkSk) mice ($n = 18$) after injection of 100 μg MCA.

Given the comparable effects on tumour development of eliminating lymphocytes or STAT1-dependent IFNγ signalling, either individually or together, there is an extensive overlap between the two tumour-suppressor systems.

To determine whether these tumour-suppressor systems also function to prevent the formation of spontaneous tumours, we monitored tumour development in unmanipulated 129/SvEv wild-type mice, RAG2$^{-/-}$ mice and RkSk mice. As neither wild-type nor RAG2$^{-/-}$ mice showed outward signs of disease during the observation period, animals were killed and evaluated for neoplasia upon reaching a minimum age of 15 months. At necropsy, 9/11 wild-type mice (15–21 months) were free of neoplastic disease, one (16.1 months) developed an intestinal adenoma and one (19 months) displayed a harderian gland cystadenoma, but none had cancer (Fig. 2a and Supplementary Information). In contrast, 12/12 RAG2$^{-/-}$ mice of ages 15–16 months displayed neoplastic lesions in the intestinal tract and elsewhere (Fig. 2b and Supplementary Information). Half of these mice developed malignant neoplasias (three had caecal adenocarcinomas, one had an adenocarcinoma at the ileo-caecal junction, one had an adenocarcinoma of the small intestine and one had a lung adenocarcinoma) that grew progressively when transplanted into naive RAG2$^{-/-}$ mice. Six of eleven RkSk mice developed overt mammary-gland carcinomas (one had two mammary adenocarcinomas and another carried distinct adenocarcinomas in the breast and caecum) that were detected well before the mice reached 15 months (Fig. 2c and Supplementary Information). This result is particularly striking given the extremely low frequency of spontaneous mammary tumours in wild-type mice or RAG2$^{-/-}$ mice of 129 origin (Fig. 2a, b) and the very late appearance (>20

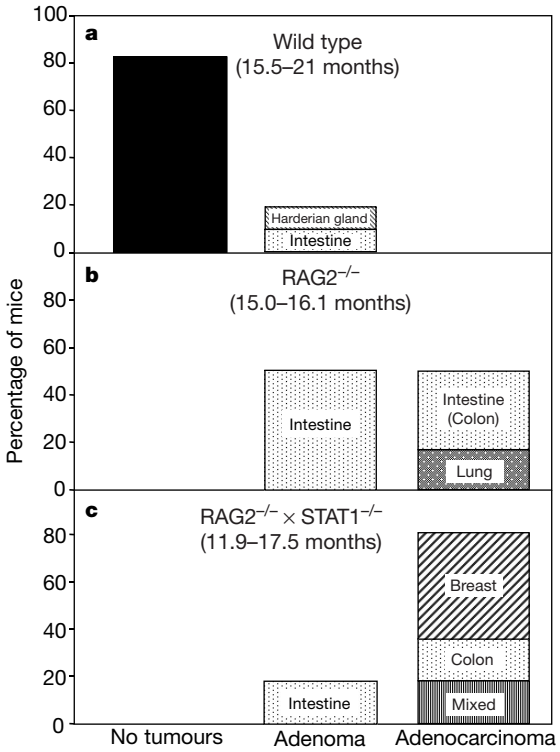


Figure 2 Increased development of spontaneous neoplastic disease in immunodeficient mice. Eleven wild-type 129/SvEv mice (**a**), 12 RAG2$^{-/-}$ mice (**b**) and 11 RkSk mice (**c**) were housed in a room that was free of specific pathogens including *Helicobacter*. Mice were killed when they appeared morbid, developed overt masses, or upon reaching an age of 15–21 months. Organs were examined for gross pathology, and histologic sections were analysed microscopically by three independent pathologists. Additional details of tumour formation and histologic sections from representative mice can be viewed as Supplementary Information.

© 2001 Macmillan Magazines Ltd

months) of mammary tumours in mice that lack only the STAT1 gene (data not shown). Cells from mammary carcinomas grew progressively when transplanted into naive RkSk mice. The other five RkSk mice did not display palpable masses but at necropsy two had adenocarcinomas in the caecum, one had adenocarcinomas in the caecum and lung and the other two displayed intestinal adenomas. In total, 82% of the RkSk group developed spontaneous cancer. Thus, mice that lack lymphocytes, either alone or in combination with an IFNγ signalling deficit, are significantly more prone to spontaneous epithelial tumour development than their wild-type counterparts ($P < 0.01$ and 0.001, respectively). Moreover, because RkSk mice develop more spontaneous cancers than RAG2$^{-/-}$ mice ($P < 0.01$), the tumour-suppressor mechanisms mediated by lymphocytes and IFNγ/STAT1 must only partially overlap.

To assess whether the immune system influences the immunogenic phenotype of tumours formed during chemical carcinogenesis, the tumorigenicities of MCA-induced sarcomas from RAG2$^{-/-}$ and wild-type mice were compared using tumour transplantation approaches. When injected into RAG2$^{-/-}$ mice, both wild-type- and RAG2$^{-/-}$-derived sarcomas grew progressively with equivalent kinetics (Fig. 3a, b). When transplanted into naive syngeneic immunocompetent hosts, 17/17 different tumours from wild-type mice grew progressively (Fig. 3c). In contrast, 8/20 (40%) distinct tumours from MCA-treated RAG2$^{-/-}$ mice were rejected following transplantation into immunocompetent mice even when injected at high tumour cell inocula (10^6 cells per mouse) (Fig. 3d). The rejection of tumours derived from the immunodeficient hosts was not due to potential minor genetic differences between wild-type 129/SvEv and RAG2$^{-/-}$ 129/SvEv mice, as the tumours were rejected when transplanted into wild-type × RAG2$^{-/-}$ F1 mice. Moreover, we ruled out the possibility that the rejection of the RAG2$^{-/-}$ tumours was due to immune responses against retrovirus-related antigens as the tumours (1) did not produce infectious amphotropic or ecotropic retrovirus (plaque assays), (2) did not express retroviral gp70 messenger RNA (RT-PCR) and (3) did not express MuLV-related antigens (flow cytometry using G_{IX} MuLV antiserum[16]) (data not shown). Thus, tumours that arise in lymphocyte-deficient mice are more immunogenic than those that develop in the presence of an intact immune system. Taken together with our previous findings that IFNγ-unresponsive mice develop more tumours than wild-type mice and that IFNγ responsiveness of the tumour cell is essential for effective immune recognition[9], these results show that lymphocytes and the IFNγ/STAT1 signalling pathway collaborate with one another to shape the immunogenic phenotype of tumours that eventually form in immunocompetent hosts.

To determine the link between the lymphocyte-dependent and IFNγ/STAT1-dependent tumour-suppressor processes, we asked whether the highly tumorigenic phenotype of IFNγ-insensitive tumour cells could be eliminated by selectively expressing in them components of the MHC class I processing and presentation pathway that are known to be significantly upregulated by IFNγ. We studied two pathway components, TAP1 and the H-2K^b heavy chain, because the decreased or absent expression of these proteins is a mechanism that tumours use to escape immune detection[17]. This mechanism is also used by tumours that develop inactivating mutations in the genes that encode the proximal IFNγ receptor signalling proteins (ref. 9; V.S. & R.D.S., unpublished work). Our focus on TAP1 was also prompted by a previous report showing that expression of rat TAP1 in a mouse small-cell lung-carcinoma cell line decreased its tumorigenicity[18], although this study was complicated by the use of xenogeneic TAP1, which differs from the endogenous mouse protein at 80 positions[19].

We chose to use for these studies two distinct IFNγ-insensitive sarcomas (RAD.gR.28 and RAD.gR.30) derived from IFNGR1$^{-/-}$ 129/SvEv mice because we had (1) established previously that

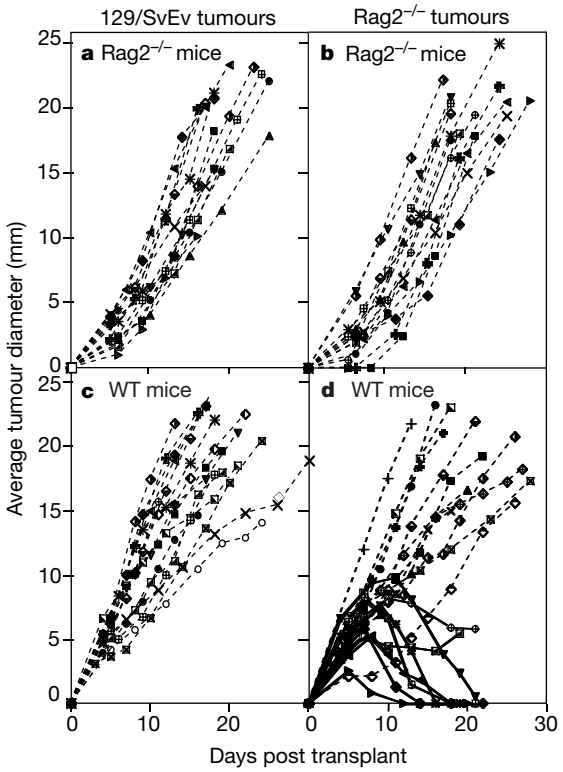


Figure 3 Increased immunogenicity of tumours derived from MCA-treated RAG2$^{-/-}$ mice. Immunodeficient RAG2$^{-/-}$ hosts were inoculated on day 0 with a dose of 10^5 tumour cells derived from wild-type 129/SvEv mice (**a**) or RAG2$^{-/-}$ mice (**b**). Tumour growth is plotted as mean tumour diameter of 3–5 mice inoculated with a distinct tumour. Groups of 5–8 immunocompetent 129/SvEv × RAG2$^{-/-}$ F1 mice were inoculated on day 0 with doses of 10^6 tumour cells derived from 17 individual 129/SvEv mice (**c**) or 20 individual RAG2$^{-/-}$ mice (**d**) and tumour growth monitored as above. In **d**, the dashed lines denote tumours that grew progressively, whereas solid lines represent tumours that were rejected.

© 2001 Macmillan Magazines Ltd

sarcomas lacking either the IFNGR1 subunit or STAT1 display common biologic response deficits and enhanced *in vivo* growth and (2) defined, in great detail, the *in vivo* growth behaviour of IFNGR1$^{-/-}$ sarcomas before and after reconstitution of IFNγ sensitivity[9]. Both tumour cell lines, which express low but detectable amounts of TAP1 and H-2K^b protein (H-2K^b), were stably transfected with expression plasmids encoding the 129/SvEv haplotypes of TAP1 or H-2K^b, and clones were selected that expressed high protein levels comparable to those expressed in IFNγ-treated, IFNγ-responsive cells (see Supplementary Information). Parental RAD.gR28 cells (Fig. 4a), empty vector-transfected RAD.gR28.neo cells (Fig. 4b) and 2/2 clones of H-2K^b-transfected RAD.gR28.K^b cells (Fig. 4a) grew progressively in immunocompetent mice when injected at 10^6 cells per mouse. In contrast, TAP1-transfected RAD.gR28.TAP1 cells formed small subcutaneous masses that expanded for the first 5–10 days but then disappeared two weeks after inoculation (Fig. 4b). This effect was generalizable to 8/8 independent RAD.gR28.TAP1 clones (two representative clones are shown in Fig. 4b). In addition, the same result was obtained with 2/2 clones of a second IFNγ-insensitive tumour cell line (RAD.gR30) engineered for TAP overexpression (Fig. 4c). Rejection of TAP1-transfected IFNγ-insensitive tumour cells required the participation of both CD4$^+$ and CD8$^+$ T cells because rejection of the TAP1-reconstituted tumour cells occurred only in wild-type mice and not in lymphocyte-deficient RAG2$^{-/-}$ mice (Fig. 4d) and was abrogated in wild-type recipients depleted of either CD4$^+$ or CD8$^+$ T cells (Fig. 4e). We found that wild-type mice that rejected the TAP1-reconstituted IFNγ-insensitive tumour were immune to subsequent challenge with either the same reconstituted tumour cell line or the unreconstituted IFNγ-insensitive parental tumour cell line (Fig. 4f), a result that supports the concept that higher antigen expression is needed for the induction of an immune response to a tumour than is required for recognition of a tumour by a pre-sensitized host. Taken together, these results reveal that TAP1 represents at least one link between tumour suppression manifested by the IFNγ signalling pathway and lymphocytes.

In this study, we show that the immune response is essential to prevent the development of carcinogen-induced sarcomas and spontaneous epithelial tumours. We also show that the tumour-suppressor function of the immune system is critically dependent on the actions of IFNγ, which, at least in part, are directed at regulating tumour-cell immunogenicity. These studies thus provide experimental evidence that, in principle, supports the original concept of cancer immunosurveillance. However, as tumours that develop in the presence of an intact immune system are less immunogenic than those that develop in immunodeficient hosts, the actions of the immune system also paradoxically favour the eventual outgrowth of tumours that are more capable of escaping immune detection. Thus, tumours are imprinted by the immunologic environment in which they form. For this reason, 'cancer immunosurveillance' is not the best term to describe the process because the concept, in its original form, implied that the immune system is involved only at the initial stages of cellular transformation and plays a purely protective role[1,2]. Rather, we favour the use of the term 'cancer immunoediting' to describe better the protective and sculpting actions of the immune response on developing tumours that probably occur continuously during tumour development. We envisage the scope of this process to be very broad, with a potential

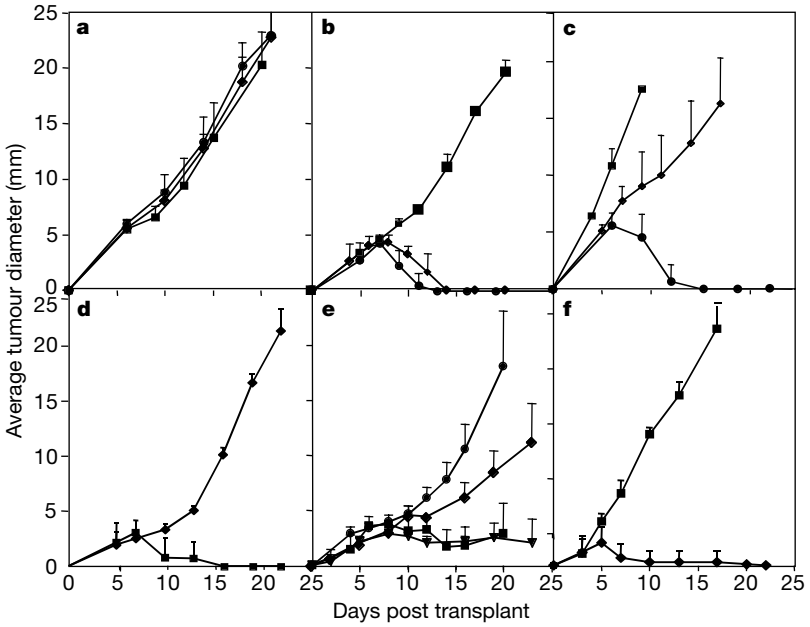


Figure 4 T-lymphocyte-dependent rejection of RAD.gR28.TAP1 cells by immunocompetent hosts. **a**, Wild-type mice were injected subcutaneously on day 0 with inocula of 10^6 RAD.gR28 (squares) cells or two representative H-2K^b-transfected RAD.gR28 clones denoted RAD.gR28.K^b.15 (diamonds) or RAD.gR28.K^b.30 (circles). **b**, One million mock-transfected RAD.gR28.neo cells (squares) or the same number of two representative TAP1-transfected RAD.gR28 clones denoted RAD.gR28.TAP1.8 (diamonds) or RAD.gR28.TAP1.16 (circles), were injected into syngeneic immunocompetent animals. **c**, 129/SvEv mice were injected with 10^6RAD.gR30 (squares), RAD.gR30.neo (diamonds) or RAD.gR30.TAP1 (circles) cells. Tumour diameter in the syngeneic hosts was monitored over time. **d**, RAD.gR28.TAP1 cells (10^6 cells per mouse) were injected into the flanks of wild-type mice (squares) or RAG2$^{-/-}$ mice (diamonds) and monitored for tumour growth. **e**, In a separate experiment RAD.gR28.TAP1 growth (10^6 cells per mouse) was compared after injection into wild-type 129/SvEv mice (squares) and wild-type mice depleted of CD8$^+$ T cells (circles) or wild-type mice (triangles) and wild-type mice depleted of CD4$^+$ T cells (diamonds). CD8$^+$ T cells were depleted by injecting mice intraperitoneally with 100 μg of purified YTS-169.4 anti-CD8 mAb seven days before tumour transplantation and once weekly thereafter. CD4$^+$ T cells were depleted by injection of 250 μg GK1.5 anti-CD4 mAb two days before tumour transplantation and every other day thereafter. Depletion of each T-cell subset was monitored by flow cytometry analysis on peripheral blood using FITC-labelled CD4 and CD8-specific mAbs directed against epitopes that were distinct from those recognized by the depleting antibodies. **f**, Growth of 10^6 parental RAD.gR28 tumour cells was assessed in naive 129/SvEv animals (squares) or 129/SvEv mice that had rejected on inoculum of 10^6 live RAD.gR28.TAP1 cells three weeks earlier (diamonds). Data are represented as average tumour diameter ± s.d. of 4–6 mice per group as a function of time.

© 2001 Macmillan Magazines Ltd

to (1) promote complete elimination of some tumours, (2) generate a non-protective immune state to others, or (3) favour the development of immunologic anergy/tolerance/indifference. Future work is needed to define the molecular basis of the cancer immunoediting process. □

Methods

Mice

RAG2$^{-/-}$, IFNGR1$^{-/-}$ (ref. 20) and wild-type mice, all on a 129/SvEv background, were either purchased from Taconic Farms or bred in our specific pathogen-free animal facility. STAT1$^{-/-}$ mice, generated in our laboratory, were maintained on a pure 129/SvEv background[14]. For the generation of RAG2$^{-/-}$ × STAT1$^{-/-}$ mice, the status of the RAG2 locus was followed by assaying blood for CD3^{+} cells by FACS as previously described[21]. Genotyping of the STAT1 locus was performed by polymerase chain reaction as previously described[9].

MCA tumour induction and transplantation

Performed as previously described[9]. A new preparation of 3′-methylcholanthrene dissolved in peanut oil was used in each experiment. Progressively growing masses 9 mm and larger were scored as tumours and confirmed by histology. The threshold value of 9 mm was chosen because masses of this size invariably continued to increase in size. To eliminate the possibility that the rejection of tumours derived from RAG2$^{-/-}$ mice was due to minor antigenic differences between RAG2$^{-/-}$ mice and 129/SvEv mice, RAG2$^{-/-}$ or Taconic 129/SvEv tumours were transplanted into immunocompetent mice that were generated by mating male 129/SvEv mice (purchased from Taconic) to female RAG2$^{-/-}$ mice.

Determination of retrovirus expression in tumour cells

The feline S^{+}L^{-} focus assay (which tests for live amphotropic retrovirus) and the XC plaque assay (which tests for live B- and N-ecotropic murine retrovirus) were performed on supernatants from 12 different RAG2$^{-/-}$ tumour cultures by BioReliance. Eighteen different sarcomas from RAG2$^{-/-}$ mice and eleven different sarcomas derived from 129/SvEv wild-type animals were tested for reactivity to the G_{IX} rat antisera that reacts with several murine leukaemia virus (MuLV) proteins[16]. Primers within conserved regions of the gp70 gene were used to amplify a 335-base-pair band from the reverse-transcribed RNA of 16 different RAG2$^{-/-}$ tumours and 7 different wild-type tumours.

Received 12 January; accepted 28 February 2001.

1. Thomas, L. in *Cellular and Humoral Aspects of Hypersensitivity States* (ed. Lawrence, H. S.) 529–532 (Hoeber, New York, 1959).
2. Burnet, F. M. The concept of immunologic surveillance. *Prog. Exp. Tumor Res.* **13,** 1–27 (1970).
3. Stutman, O. Spontaneous tumors in nude mice: effect of the viable yellow gene. *Exp. Cell. Biol.* **47,** 129–135 (1979).
4. Stutman, O. Tumor development after 3-methylcholanthrene in immunologically deficient athymic-nude mice. *Science* **183,** 534–536 (1970).
5. Stutman, O. Chemical carcinogenesis in nude mice: comparison between nude mice from homozygous and heterozygous matings and effect of age and carcinogen dose. *J. Natl Cancer Inst.* **2,** 353–358 (1979).
6. Hunig, T. T-cell function and specificity in athymic mice. *Immunol. Today* **4,** 84–87 (1983).
7. Maleckar, J. R. & Sherman, L. A. The composition of the T cell receptor repertoire in nude mice. *J. Immunol.* **138,** 3873–3876 (1987).
8. Dighe, A. S., Richards, E., Old, L. J. & Schreiber, R. D. Enhanced in vivo growth and resistance to rejection of tumor cells expressing dominant negative IFNγ receptors. *Immunity* **1,** 447–456 (1994).
9. Kaplan, D. H. *et al.* Demonstration of an IFNγ dependent tumor surveillance system in immunocompetent mice. *Proc. Natl Acad. Sci. USA* **95,** 7556–7561 (1998).
10. van den Broek, M. F. *et al.* Decreased tumor surveillance in perforin-deficient mice. *J. Exp. Med.* **184,** 1781–1790 (1996).
11. Smyth, M. J. *et al.* Differential tumor surveillance by natural killer (NK) and NKT cells. *J. Exp. Med.* **191,** 661–668 (2000).
12. Smyth, M. J. *et al.* Perforin-mediated cytotoxicity is critical for surveillance of spontaneous lymphoma. *J. Exp. Med.* **192,** 755–760 (2000).
13. Shinkai, Y. *et al.* RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *Cell* **68,** 855–867 (1992).
14. Meraz, M. A. *et al.* Targeted disruption of the STAT1 gene in mice reveals unexpected physiologic specificity in the Jak-STAT signaling pathway. *Cell* **84,** 431–442 (1996).
15. Durbin, J. E., Hackenmiller, R., Simon, M. C. & Levy, D. E. Targeted disruption of the mouse *Stat1* gene results in compromised innate immunity to viral infection. *Cell* **84,** 443–450 (1996).
16. Old, L. J. & Stockert, E. Immunogenetics of cell surface antigens of mouse leukemia. *Annu. Rev. Genet.* **17,** 127–160 (1977).
17. Marincola, F. M., Jaffee, E. M., Hicklin, D. J. & Ferrone, S. Escape of human solid tumors from T-cell recognition: molecular mechanisms and functional significance. *Adv. Immunol.* **74,** 181–273 (2000).
18. Alimonti, J. *et al.* TAP expression provides a general method for improving the recognition of malignant cells in vivo. *Nature Biotechnol.* **18,** 515–520 (2000).
19. Marusina, K., Iyer, M. & Monaco, J. J. Allelic variation in the mouse Tap-1 and Tap-2 transporter genes. *J. Immunol.* **158,** 5251–5256 (1997).
20. Huang, S. *et al.* Immune response in mice that lack the interferon-γ receptor. *Science* **259,** 1742–1745 (1993).
21. Dighe, A. S. *et al.* Tissue specific targeting of cytokine unresponsiveness in transgenic mice. *Immunity* **3,** 657–666 (1995).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank J. J. Monaco for providing a cDNA encoding TAP1 and T. H. Hansen for providing TAP1 antibodies and the cDNA encoding H-2K^{b}. We thank M. La Regina for performing the veterinary pathology and C. Hughley and C. Arthur for expert technical assistance. We also wish to thank P. M. Allen, K. M. Murphy, A. S. Shaw and E. R. Unanue for helpful suggestions during the preparation of this manuscript and J. F. Piccirillo for assistance with the statistical analysis of the data. This work was supported by grants from the NIH and the Cancer Research Institute. V.S. was supported and A.T.B. is supported by a pre-doctoral training grant from the Cancer Research Institute (CRI) and H.I. is supported by the John Hans and Edna Alice Old Post-Doctoral Fellowship from the CRI.

Correspondence and requests for materials should be addressed to R.D.S. (e-mail: schreiber@immunology.wustl.edu).

Somatic activation of the *K-ras* oncogene causes early onset lung cancer in mice

Leisa Johnson*†, Kim Mercer*‡, Doron Greenbaum*§, Roderick T. Bronson‖, Denise Crowley*‡, David A. Tuveson*‡¶ & Tyler Jacks*‡

* *Department of Biology, Massachusetts Institute of Technology, and ‡ Howard Hughes Medical Institute, Center for Cancer Research, Cambridge, Massachusetts 02139, USA*

‖ *Department of Pathology, Tufts University Schools of Medicine and Veterinary Medicine, Boston, Massachusetts 02111, USA*

¶ *Dana-Farber Cancer Institute, Division of Adult Oncology, Boston, Massachusetts 02115, USA*

About 30% of human tumours carry *ras* gene mutations[1,2]. Of the three genes in this family (composed of *K-ras*, *N-ras* and *H-ras*), *K-ras* is the most frequently mutated member in human tumours, including adenocarcinomas of the pancreas (~70–90% incidence), colon (~50%) and lung (~25–50%)[1–6]. To constuct mouse tumour models involving *K-ras*, we used a new gene targeting procedure to create mouse strains carrying oncogenic alleles of *K-ras* that can be activated only on a spontaneous recombination event in the whole animal. Here we show that mice carrying these mutations were highly predisposed to a range of tumour types, predominantly early onset lung cancer. This model was further characterized by examining the effects of germline mutations in the tumour suppressor gene *p53*, which is known to be mutated along with *K-ras* in human tumours. This approach has several advantages over traditional transgenic strategies, including that it more closely recapitulates spontaneous oncogene activation as seen in human cancers.

The effects of *ras* gene mutations have been studied in transgenic mice; however, most strains expressed *H-ras* or *N-ras* (reviewed in ref. 7). Traditional transgenic strategies direct expression of the oncogene in all cells of the target tissue and may lead to supra-physiological levels of expression. In an effort to construct a *ras*-based mouse tumour model that overcomes these limitations, we have used a variation of 'hit-and-run' gene targeting[8] to create new mouse strains harbouring latent, oncogenic alleles of *K-ras* capable of spontaneous activation *in vivo*.

Present addresses: † Onyx Pharmaceuticals, Richmond, California 94806, USA (L.J.); § Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94143, USA (D.G.).

© 2001 Macmillan Magazines Ltd

The hit-and-run gene targeting procedure has been used to introduce subtle mutations and typically involves two distinct steps of homologous recombination performed in embryonic stem (ES) cells in culture[8]. In our protocol, however, we produced animals carrying the targeted insertion allele that was created in cultured cells ('hit' step) and allowed the excision ('run') step to occur *in vivo* (see Fig. 1, Methods, and Supplementary Information). After insertional targeting of a plasmid carrying the *K-ras* exon 1 with an activating glycine to aspartic acid mutation at codon 12 (G12D), ES cell clones representing two different *K-ras* alleles were identified and used for blastocyst injection and germline transmission (Fig. 1). Hereafter, we refer to these two alleles as $K\text{-}ras^{LA1}$ and $K\text{-}ras^{LA2}$, for latent allele types 1 and 2. These alleles differ in that $K\text{-}ras^{LA2}$ has two mutant copies of exon 1, whereas in $K\text{-}ras^{LA1}$, only the upstream copy of exon 1 is mutant (Fig. 1). In $K\text{-}ras^{LA1}$, half of the subsequent *in vivo* recombination events would produce an active $K\text{-}ras^{G12D}$ allele and half would produce wild-type *K-ras*, whereas in $K\text{-}ras^{LA2}$, all such events would generate the oncogenic form of the gene (Fig. 1; and Supplementary Information). Germline chimaeras from two independent clones of both $K\text{-}ras^{LA1}$ and $K\text{-}ras^{LA2}$ were bred to C57Bl/6 females to create C57Bl/6/129/sv F_1 mutants that we monitored over time for signs of disease; the mutations were also studied on a 129/sv inbred background (see Supplementary Information). Because the $K\text{-}ras^{LA}$ alleles were expected to be non-functional in the germline configuration, they were maintained in a heterozygous state (as *K-ras* is an essential gene in the mouse[9]). The recombination frequency of duplicated genomic sequences, which can occur through intrachromosomal recombination or unequal sister-chromatid exchange, has been estimated to range between 10^{-3} and 10^{-7} per cell generation, depending on the locus[8,10]. Recombination frequencies in this range would ensure that animals carrying the $K\text{-}ras^{LA}$ alleles would have many, presumably widely distributed, cells expressing the *K-ras* oncogene.

As shown in Fig. 2a, both $K\text{-}ras^{LA1}$ and $K\text{-}ras^{LA2}$ mutations caused significantly reduced survival compared with wild-type controls, with a mean age of death/sacrifice of around 300 days for the $K\text{-}ras^{LA1}$ strain and 200 days for the $K\text{-}ras^{LA2}$ strain. At necropsy, all animals were found to have an extensive tumour burden, and, as anticipated, the $K\text{-}ras^{LA2}$ strain developed a more rapid tumour phenotype. As summarized in Fig. 2b, the $K\text{-}ras^{LA1}$ and $K\text{-}ras^{LA2}$ mutant strains developed a similar spectrum of tumours. The most frequent site of tumour occurrence was in the lung, where 100% of animals developed multifocal tumours at different stages of progression. These tumours were first detectable as small pleural nodules at one week of age (Figs 2c and 3a). Lungs from one-day-old pups did not contain any tumours or pre-neoplastic lesions, consistent with the finding that *K-ras* expression increases

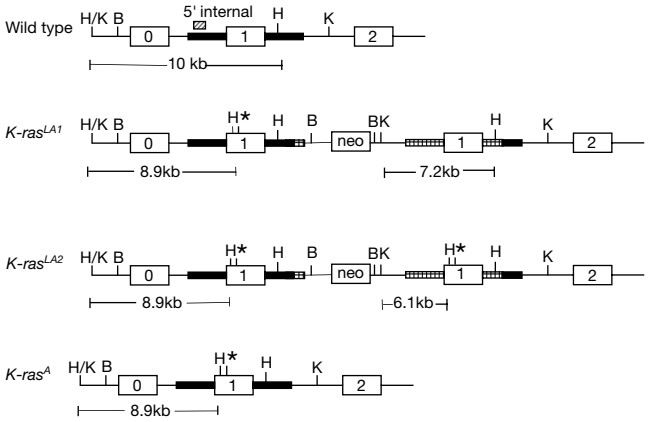


Figure 1 *K-ras* alleles. Gene targeting created two $K\text{-}ras^{LA}$ alleles in ES cells (see Supplementary Information for details). The $K\text{-}ras^{LA1}$ allele carries the Asp 12 mutation solely in the 5′ copy of exon 1, whereas both copies of exon 1 carry the mutation in the $K\text{-}ras^{LA2}$ allele. $K\text{-}ras^{A}$ (active) allele is detected in tumours from $K\text{-}ras^{LA}$mice. Although the $K\text{-}ras^{LA1}$ allele can also produce a wild-type allele on recombination, tumours will contain the mutant Asp 12 allele. Asterisk, Asp 12 mutation; H, *Hin*dIII; K, *Kpn*1; B, *Bam*H1; *neo*, neomycin gene. The targeting vector is hatched and the exons are numbered. Relevant probes and restriction lengths are shown.

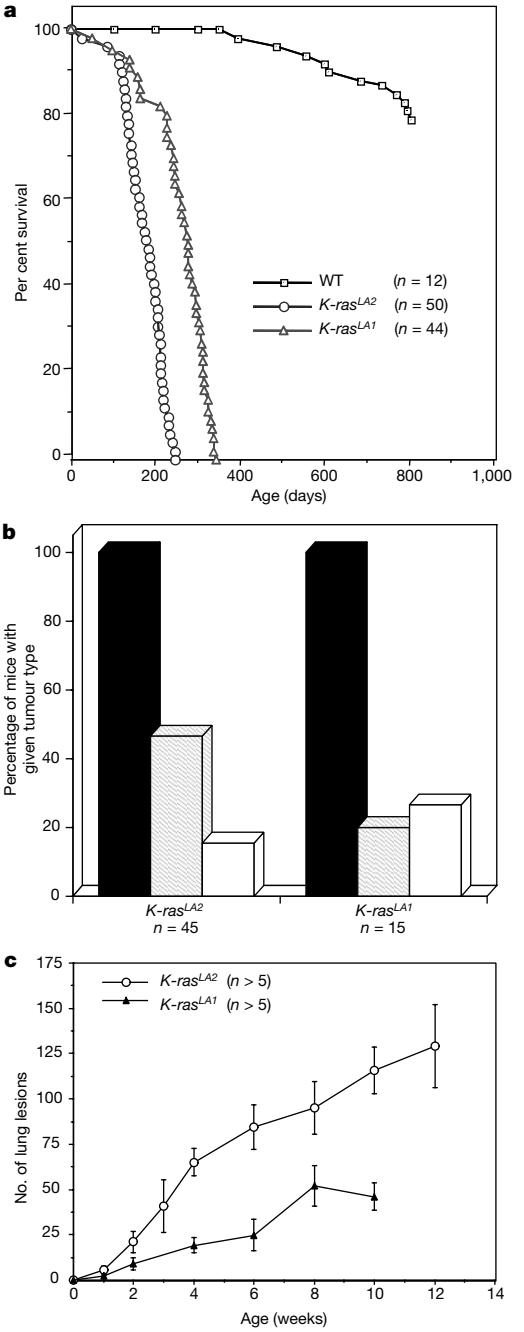


Figure 2 Effect of $K\text{-}ras^{LA}$ mutations on survival and tumour incidence. **a**, $K\text{-}ras^{LA}$C57Bl6/129/sv F_1 mutant mice have significantly reduced life spans compared with wild-type controls, with $K\text{-}ras^{LA2}$ mice having decreased survival compared with $K\text{-}ras^{LA1}$ mice. **b**, Tumour spectra in $K\text{-}ras^{LA1}$ and $K\text{-}ras^{LA2}$C57Bl6/129/sv F_1 mice. Lung adenocarcinoma, thymic lymphoma and papilloma tumours are indicated by filled, hatched and empty columns, respectively. **c**, Lung tumour incidence and burden. Mice were killed and examined for the presence of visible lesions on the pleural surface of their lungs. Lesions (mean ± s.d.) increased in both number and size with age.As expected, $K\text{-}ras^{LA2}$ mutant mice developed more lesions than their $K\text{-}ras^{LA1}$ counterparts.

© 2001 Macmillan Magazines Ltd

significantly in rodent lungs shortly after birth[11]. Tumour multiplicity and size increased with age (Figs 2c and 3a, b), ultimately resulting in respiratory distress and death or sacrifice of the animal.

With regards to histopathology, *K-ras*^{LA} lung tumours seem to evolve through a series of morphological stages from mild hyperplasia/dysplasia to overt carcinoma, reminiscent of human non-small cell lung cancer (NSCLC). The smallest lesions consisted of hyperplastic alveolar epithelium (Fig. 3c), and closely resembled human 'atypical adenomatous hyperplasia', a proposed precursor dysplastic lesion to invasive lung carcinoma[12,13]. As the lesions enlarged, they formed small tumours termed alveolar adenomas, composed of a monomorphous population of airway epithelium with minimal cytologic atypia (Fig. 3d). Some alveolar adenomas showed areas of glandular differentiation (Fig. 3e) and papillary architecture (Fig. 3f). About 5–10% of the tumours showed features of well differentiated human papillary adenocarcinoma, including nuclear enlargement, prominent nucleoli and increased mitotic rate (Fig. 3g). In older mice, we observed infrequent metastases of the *K-ras*^{LA} lung tumours to thoracic lymph nodes, the kidney and other visceral organs (Fig. 3i). Immunohistochemical analysis demonstrated that the lung tumours expressed surfactant apoprotein-C and -A, but not Clara cell antigen (Fig. 3j, k; and data not shown), indicating an alveolar type II cell lineage. Human lung adenocarcinomas also frequently express markers of alveolar type II cells[14].

In addition to lung cancer, the *K-ras*^{LA} animals were also prone to both thymic lymphoma and skin papillomas, each occurring in about 30% of the animals collectively (Fig. 2b; and Supplementary Information). The papillomas typically arose in areas subject to abrasion (that is, the ears and snout), were predominantly

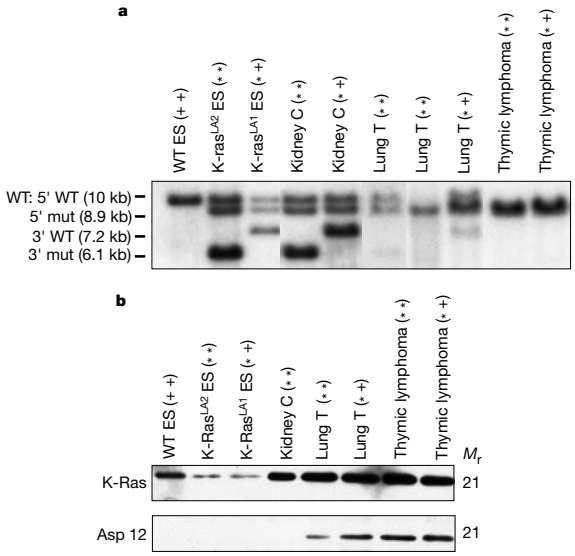


Figure 4 Evidence of recombination in *K-ras*^{LA} tumours. **a**, **b**, DNA (**a**) and protein (**b**) were prepared from tumours (T) and normal tissues (C). Control samples were obtained from wild-type (WT), *K-ras*^{LA2} (**) and *K-ras*^{LA1} (*+) ES cells. **a**, Southern blot of *Hin*dIII–*Kpn* 1-digested DNA that was probed with a 5′ internal probe demonstrates recombination in tumours (see Fig. 1). Loss of the wild-type *K-ras* allele and/or amplification of the mutant allele is shown in some tumours. **b**, The rearranged *K-ras*^{LA} alleles are expressed in tumours. Extracts were immunoprecipitated with a pan-Ras antibody, followed by immunoblotting with antibodies specific to wild-type K-Ras and mutant K-Ras^{G12D}.

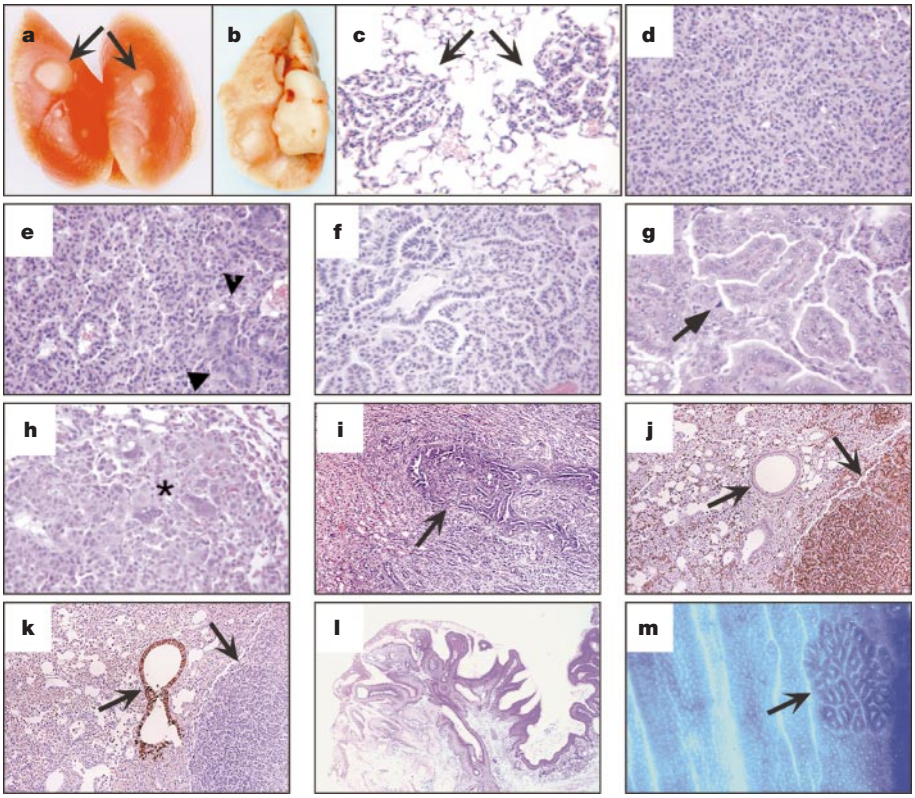


Figure 3 Tumour pathology in *K-ras*^{LA}mutant mice. **a**, Thirty-day-old mouse lungs showing numerous pleural lesions (arrows). **b**, Advanced lung tumours in the lungs of a 150-day-old mouse. **c**, Alveolar adenomatous hyperplasia 2 weeks after birth. **d**, Lung alveolar adenoma. **e,** Alveolar adenoma showing focal glandular differentiation (arrowheads). **f**, Papillary adenoma. **g**, Well-differentiated lung papillary adenocarcinoma displaying mitosis (arrow). **h**, Poorly differentiated lung tumour from a *K-ras*^{LA}; *p53*$^{-/-}$ mouse showing tumour giant cells (asterisk). **i**, Renal metastasis. **j**, Pro-SP-C-positive lung adenoma (right arrow) and Pro-SP-C-negative bronchiole (left arrow). **k**, CCA-negative lung tumours (right arrow) and CCA-positive terminal bronchioles (leftarrow). **l**, Oral cavity skin papilloma. **m**, Colonic aberrant crypt foci (ACF) (arrow).

© 2001 Macmillan Magazines Ltd

pedunculated (Fig. 3l), and demonstrated limited, if any, progression to squamous cell carcinomas during the lifespan of these animals. The frequency of lymphoma and papilloma differed depending on genetic background (see Supplementary Information).

Despite the frequency of *K-ras* mutations in carcinomas of the pancreas and colon of humans, we did not detect these tumours in the *K-ras*^{LA} mice ($n = 149$). Significantly, however, all of the mutant mice examined ($n = 21$) had multiple aberrant crypt foci (ACF) of the colon (Fig. 3m; and Supplementary Information), whereas wild-type littermates had none. ACF are often observed in patients with colon cancer, and are microscopic lesions consisting of clusters of abnormally large crypts with an irregularly shaped lumen, increased pericryptal zone and elevated thick epithelium[15].

Evidence for recombination of the latent allele in tumours was obtained through the analysis of DNA, messenger RNA, and protein (Fig. 4; and Supplementary Information). Southern blot analysis showed loss of one duplicate copy of exon 1 from the *K-ras*^{LA} allele in tumour DNA (Fig. 4a). Some of the lung tumour samples contained residual signal from the latent allele, probably owing to macrophage infiltration (Fig. 4a; and data not shown). Notably, 50% of all thymic lymphomas examined ($n = 33$), as well as some of the lung tumours, showed a greater than 1/1 ratio of mutant/wild-type alleles (Fig. 4a; 8.9-kilobase (kb) and 10-kb bands, respectively). Furthermore, in four of the five lung tumours tested, the mutant *K-ras* mRNA was expressed at higher levels than the wild-type allele (data not shown). This may reflect amplification of the mutant allele and/or loss of the wild-type allele in a fraction of tumour cells, as has been demonstrated for *ras* oncogenes in human and murine tumours[3,16–19]. Laser capture microdissection[20] and polymerase chain reaction (PCR) was used to show that recombination had occurred within the tumours (data not shown). The expression of the mutant *K-ras* allele was demonstrated by PCR with reverse transcription (RT–PCR) analysis of mRNA (see Supplementary Information) and immunoblotting using polyclonal antibodies specific to the K-Ras^{G12D} protein (Fig. 4b).

In human adenocarcinomas of the lung, colon and pancreas, *K-ras* mutations are frequently associated with alterations in the *p53* pathway[5,6,21]. In addition, ectopic expression of oncogenic *ras* can cause *p53*-dependent growth arrest and apoptosis in cell-culture systems[22,23]. We therefore crossed the *K-ras*^{LA1} strain to a strain carrying a germline deletion in *p53* (ref. 24) and examined mice with the genotypes *K-ras*$^{LA1/+}$; *p53*$^{+/-}$ and *K-ras*$^{LA1/+}$; *p53*$^{-/-}$. As shown in Fig. 5, compound *p53* and *K-ras*^{LA1}mutant mice had reduced lifespans compared with those of either single mutant parental strain. Notably, lung tumours in the *K-ras*$^{LA/p53}$ double mutants showed more malignant features than the *K-ras*^{LA} tumours, including marked cellular pleomorphism and anaplasitc changes with giant-cell formation (Fig. 3h). Furthermore, mice with the genotype *K-ras*$^{LA1/+}$; *p53*$^{-/-}$ had a broader tumour spectrum compared with the parental strains, with about 30% of the double-mutant animals developing haemangiosarcomas or fibrosarcomas (Table 1).

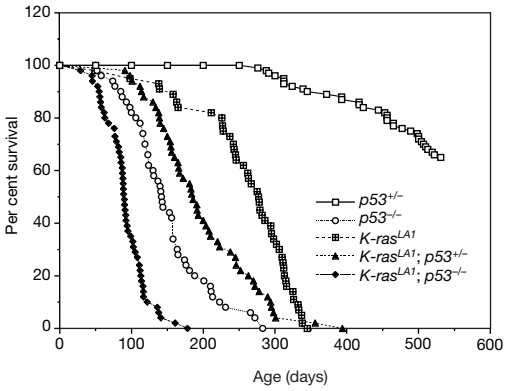


Figure 5 Cooperation between *K-ras*^{LA} and *p53* mutations. The combination of *K-ras*^{LA} with a *p53* loss-of-function mutation accelerated the onset of cancer, resulting in statistically significant decreased survival as compared with either mutation in isolation.

Here we describe a new, general strategy for constructing cancer-prone mouse strains and, in particular, the development of strains with profound predisposition to lung cancer resembling NSCLC in humans. Through an adaptation of the hit-and-run gene targeting protocol, we have created a situation in which multiple cells of the lung and other tissues of the mouse independently acquire oncogenic *K-ras* mutations and, thereby, initiate neoplastic progression. We have also demonstrated that mutation of *p53* cooperates in the progression of lung tumours in *K-ras*^{LA} mice, furthering the relevance to human lung cancer development. Future studies will focus on determining what other germline mutations can cooperate with the *K-ras*^{LA} mutation to accelerate (or inhibit) tumour development in the lung as well as to characterize what mutations occur spontaneously in this model. Of note, preliminary analysis of tumour DNA and protein indicates that *p16*^{Ink4A}, which is deleted or epigenetically silenced in up to 60% of human NSCLC (ref. 21), is not a frequent mutational target in this model (J. Sage, unpublished observations).

There are two general explanations to account for the tumour spectrum in the *K-ras*^{LA} mouse strains. First, the activating rearrangement event may occur at varying frequencies in different tissues of the mouse, a possibility that we are unable to address experimentally with available reagents. Alternatively, cells of the lung (as well as thymus and skin) may be especially sensitive to the proliferative effects of oncogenic *K-ras*, such that a tumour will more readily arise when the recombination event occurs in such cell types. Indeed, *K-ras* mutations occur frequently in both spontaneous and chemically induced lung tumours in mice (reviewed in ref. 25). In other cell types, expression of oncogenic *K-ras* may have no effect on its own or may produce an alternative cellular phenotype, such as abnormal differentiation, cell cycle arrest, senescence or apoptosis.

The possibility of cell-type-specific responses to oncogenic *K-ras* expression is strongly supported by the development of non-progressing ACF in *K-ras*^{LA} mice. In humans, *K-ras* mutations have also been linked to the development of benign ACF; however, in the context of a pre-existing *APC* gene mutation, *K-ras* mutations are associated with tumour progression[26]. Thus, the relative order of *ras* gene mutations may also be important in determining their potential tumorigenic effects. In this regard, we have crossed the *K-ras*^{LA} strains with the *Apc*^{min} strain[27] to assess a possible genetic interaction between these two genes in the mouse. However, we failed to observe a change in the intestinal polyp phenotype associated with *Apc*^{min} (data not shown), possibly owing to the relatively low frequency of the *K-ras* rearrangement event given the number of animals examined ($n = 54$).

Because of its frequent involvement in human cancer, the Ras pathway is considered an attractive target for chemotherapeutic

Table 1 Effect of *p53* status on the tumour spectrum in *K-ras*^{LA} mice

Tumour type	*K-ras*$^{LA1+/-}$; *p53*$^{+/-}$	*K-ras*$^{LA1+/-}$; *p53*$^{-/-}$
Lung adenocarcinoma	100	100
Thymic lymphoma	37	39
Papilloma	24	4
Fibrosarcoma	3	28
Haemangiosarcoma	2	33
Duodenal adenocarcinoma	0	2
Undifferentiated sarcoma	1	0
Histiocytic sarcoma	1	4
Medullablastoma	0	2
Osteosarcoma	0	2
Teratoma	0	2

Percentage of animals with a given tumour type. Numbers were combined for animals from both *K-ras*^{LA1} and *K-ras*^{LA2}. *K-ras*$^{LA1+/-}$; *p53*$^{+/-}$ ($n = 54$); *K-ras*$^{LA1+/-}$; *p53*$^{-/-}$ ($n = 87$).

© 2001 Macmillan Magazines Ltd

intervention[28]. The *K-ras*^{LA} mouse strains may be useful in assessing the efficacy of Ras pathway-directed therapies, especially given potential differences in the tumorigenic consequences of *K-ras* mutations versus other *ras* gene mutations, as well as the relative resistance of K-Ras to farnesyl-transferase inhibitors (reviewed in refs 29, 30). The spontaneous nature of *K-ras* activation in this model may more accurately mimic the interaction of tumour cells and their normal environment, which may be relevant to the activity and efficacy of chemotherapeutic agents. Because the tumour phenotype of the *K-ras*^{LA} mice exhibits such short latency and high penetrance, these strains could also be useful in screening potential chemopreventative agents. □

Methods

Construction of targeting vector

We used the *K-ras* genomic clone to construct the targeting vector as described[9]. A 1.9-kb *Sal*I–*Xho*I fragment containing exon 1 sequences was isolated and subcloned into pGEM-7Z$^+$ (Promega). Site-directed mutagenesis was performed using the Amersham Sculptor kit to create p*K-ras*–Ex1^{D12}. The sequence of the oligonucleotide used to create the *Hin*dIII and Asp 12 mutations was 5′-ATGACTGAGTATAA**G**CTTGTGGTGGTTGGAGCTGA**T** GGCGTAGGC-3′ (the two nucleotide alterations are indicated in bold). All clones were sequenced to insure that only the desired mutations were incorporated into exon 1. Next, p-3′-*K-ras*^{D12} was created by ligating the following fragments: a 1.9-kb *Sal*I–*Xho*I fragment from p*K-ras*–Ex1^{D12}; a 1.7-kb *Xho*I–*Xba*I fragment from p*K-ras*-3'; and a 3.0-kb *Sal*I–XhoI fragment from pKSII$^+$(Stratagene). A vector carrying the desired selectable marker was first generated by subcloning a 1.8-kb *Xho*I–*Sac*I fragment from pPGKRN containing a wild-type *neo* gene into a 5.5-kb *Xho*I–*Xba*I fragment from pPNT to create pPRNT. Next, pKSII$^+$-R*neo* was created by subcloning a 1.9-kb *Eco*RI–*Acc*65I *neo* fragment from pPRNT into *Eco*RI–*Acc*65I-digested pKSII$^+$. The final targeting construct, p*K-ras*$^{D12\text{-}LA}$, was created by ligating a 2.8-kb *Not*I–*Sal*I fragment from p*K-ras*-5′, a 3.6-kb *Sal*I–*Xba*I fragment from p-3′-*K-ras*^{D12} and a 4.8-kb *Xba*I–*Not*I fragment from pKSII$^+$–R*neo*.Exon 1 sequences in the final targeting vector were confirmed by sequencing.

Targeting and characterization of ES cell clones

D3 ES cells were cultured, electroporated, and selected as described[9]. To identify targeted clones, DNAs were digested with *Bam*HI + *Kp*nI, resolved on 0.8% agarose gels, and Southern blot analysis using a 5′ external probe was performed as described[9]. Targeted clones were analysed further using Southern blotting to determine the integration pattern and copy number. We performed sequence analysis to confirm the status of the Asp 12 mutation in each of the correctly targeted events. We used PCR to amplify *K-ras* exon 1 sequences. The 5′primer (5′-GGGTAGGTGTTGGGATAGCTGTCGACAAGC-3′) was located in intron 0 and the 3′ primer (5′-CCTTTACAAGCGCACGCAGACTGTAGAGC-3′) was located in intron 1. The 520-bp fragment was phenol-chloroform-extracted, purified in a Microcon 100 microconcentrator (Amicon), and then sequenced (US Biochemical) using a primer (5′-TCTTGTGTGAGACATG-3′) located immediately 5′ to exon 1. All ES clones were maintained in the presence of G418 to prevent the growth of cells that had undergone recombination before *in vivo* analysis.

Generation of *K-ras*^{LA} mice

C57BL/6 blastocyst-stage embryos were injected with 10–15 *K-ras*^{LA} ES cells and subsequently transferred to pseudopregnant Swiss Webster females for further development. Chimaeric mice were mated to C57BL/6 and 129/Sv animals and F_1 agouti offspring were genotyped. Germline transmission of the mutant allele was detected by either Southern blot or PCR analysis of tail DNA obtained at weaning. PCR protocols for genotyping are available on request.

RT–PCR/oligonucleotide Southern analysis

Processed mRNA (Poly (A)$^+$ RNA) was isolated from ES cells or tissue using Qiagen's Oligotex Direct mRNA kit as per the manufacturer's recommendations. Reverse transcription was performed at 70 °C for 15 min using Perkin Elmer's Reverse Transcription RNA PCR kit. Briefly, 250 ng of Poly (A)$^+$ RNA was reversed transcribed using recombinant thermostable reverse transcriptase (r*Tth*) DNA polymerase in the presence of 10 mM Tris-HCl, pH 8.3, 90 mM KCl, 1 mM $MnCl_2$, 200 μM each deoxynucleotide triphosphate, and the 3′ primer, K-Ras Ex.3-3′A: 5′-ACGGAATCCCGTAACTC-3′. Complementary DNA was purified over Qiagen Qiaquick columns, eluted with 10 mM Tris-HCl, pH 8.8, and then amplified by PCR using Clontech's Klentaq/GC melt kit according to the manufacturer's recommendations. The PCR primers used were K-Ras Ex.3-3′A and K-Ras Ex.0-5′B (5′-CATTTCGGACCCGGAGCGAGC-3′), with an annealing temperature of 60 °C and 40 rounds of amplification. PCR products were isolated on 2.5% agarose gels and transferred onto Hybond (Amersham Life Science) for oligonucleotide-specific Southern analysis. Filters were hybridized overnight at 37 °C in 5× Denhardts, 5× SSPE, 0.5% SDS and 100 mM sodium pyrophosphate, pH 7.5 containing 200 ng of the oligonucleotide probe labelled with ^{32}P at the 5′ end. The oligonucleotide probes used were as follows: wild type (5′-GGAGCTGGTGGCGTAGGCAA-3′) and Asp 12 (5′-GGAGCTGATGGCGTAGGCAA-3′). Filters were washed in 6× SSC at room temperature, followed by successive washes in 3 M tetramethylammonium chloride, 50 mM Tris-HCl, pH 8.0, 2 mM EDTA, and 0.1% SDS at room temperature and then at 61 °C.

Histological analysis and immunohistochemistry

All animals showing obvious tumours or other signs of distress were killed and subjected to full necropsy. For histological analysis, all tissues and tumours were fixed in either Bouin's fixative or 10% neutral buffered formalin, paraffin processed, sectioned at 4 μm, and stained with haematoxylin and eosin. Immunohistochemistry was performed using the Vectastain ABC kit (Vector) on 4-μm cut paraffin sections. Briefly, endogenous peroxidase was quenched using 3% H_2O_2 in distilled water. Sections were blocked for 2 h at room temerature in PBS containing 0.2% Triton X-100 and normal goat serum. They were then incubated overnight at 4°C in PBS supplemented with 0.2% Triton X-100 and the appropriate primary antibody. The following primary antibodies and their respective dilutions were used: anti-pro-SP-C at 1:750; anti-SP-A at 1:1,000; and anti-CCA at 1:1,000. The next day, sections were allowed to equilibrate to room temperature for 1 h and were then extensively washed with 0.2% Triton X-100 in PBS. All subsequent steps were followed as per the manufacturer's recommendations.

Laser capture microdissection

According to the NIH guidelines, sections were cut at 8 μm, dried at 60 °C for 15 min and stained with haematoxylin and eosin. After laser capture using the Arcturus PixCell Laser capture microdissection instrument, DNA was extracted by lysis in 30 μl of LCM buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 1% Tween-20, and 0.04% Proteinase K) at 37 °C overnight. We then heat-inactivated Proteinase K at 85 °C for 10 min. One to five microlitres of this was then used as PCR template.

K-Ras Asp 12 peptide and antibody generation

Wild-type and Asp 12 K-Ras peptides (residues 5–17, SynPep) were coupled to keyhole limpet haemocyanin and used to produce high-titre, antigen-specific, rabbit polyclonal antibodies (Pocono Rabbit Farms and Lab). Affinity purified or IgG fractions detected wild-type and K-Ras^{G12D} by immunoblotting, but were incapable of detecting K-Ras immunohistochemically.

Immunoprecipitation and western analysis

Cell lysates were prepared, immunoprecipitated with Y13-259 (Santa Cruz Biotechnology), and analysed by western blotting as described[9]. The anti-K-Ras^{D12} and anti-K-Ras^{G12} polyclonal rabbit sera were each used at a 1:1,000 dilution, whereas the F234 anti-K-Ras mouse monoclonal antibody (Santa Cruz Biotechnology) was used at 1:200. Secondary goat anti-rabbit or goat anti-mouse antibodies were both used at 1:7,000 dilution.

Received 14 November 2000; accepted 23 January 2001.

1. Bos, J. L. ras oncogenes in human cancer: a review. *Cancer Res.* **49,** 4682–4689 (1989); erratum *Cancer Res* **50,** 1352 (1990).
2. Khosravi-Far, R. & Der, C. J. The Ras signal transduction pathway. *Cancer Met. Rev.* **13,** 67–89 (1994).
3. Bos, J. L. *et al.* Prevalence of ras gene mutations in human colorectal cancers. *Nature* **327,** 293–297 (1987).
4. Mills, N. E., Fishman, C. L., Rom, W. N., Dubin, N. & Jacobson, D. R. Increased prevalence of K-ras oncogene mutations in lung adenocarcinoma. *Cancer Res.* **55,** 1444–1447 (1995).
5. Pellegata, N. S. *et al.* K-ras and p53 gene mutations in pancreatic cancer: ductal and nonductal tumors progress through different genetic lesions. *Cancer Res.* **54,** 1556–1560 (1994).
6. Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.* **319,** 525–532 (1988).
7. Adams, J. M. & Cory, S. Transgenic models of tumor development. *Science* **254,** 1161–1167 (1991).
8. Hasty, P., Ramirez-Solis, R., Krumlauf, R. & Bradley, A. Introduction of a subtle mutation into the Hox-2.6 locus in embryonic stem cells. *Nature* **350,** 243–246 (1991); erratum *Nature* **53,** 94 (1991).
9. Johnson, L. *et al.* K-ras is an essential gene in the mouse with partial functional overlap with N-ras. *Genes Dev.* **11,** 2468–2481 (1997); erratum *Genes Dev.* **11,** 3277 (1997).
10. Seperack, P. K., Strobel, M. C., Corrow, D. J., Jenkins, N. A. & Copeland, N. G. Somatic and germ-line reverse mutation rates of the retrovirus-induced dilute coat-color mutation of DBA mice. *Proc. Natl Acad. Sci. USA* **85,** 189–192 (1988).
11. Thrane, E. V. *et al.* Differential distribution and increased levels of ras proteins during lung development. *Exp. Lung Res.* **23,** 35–49 (1997).
12. Nakanishi, K. Alveolar epithelial hyperplasia and adenocarcinoma of the lung. *Arch. Pathol. Lab. Med.* **114,** 363–368 (1990).
13. Kitamura, H., Kameda, Y., Ito, T. & Hayashi, H. Atypical adenomatous hyperplasia of the lung. Implications for the pathogenesis of peripheral lung adenocarcinoma. *Am. J. Clin. Pathol.* **111,** 610–622 (1999).
14. Linnoila, R. I., Mulshine, J. L., Steinberg, S. M. & Gazdar, A. F. Expression of surfactant-associated protein in non-small-cell lung cancer: a discriminant between biologic subsets. *J. Natl Cancer Inst. Monogr.* **13,** 61–66 (1992).
15. Pretlow, T. P. *et al.* Aberrant crypts: putative preneoplastic foci in human colonic mucosa. *Cancer Res.* **51,** 1564–1567 (1991).
16. Bos, J. L. The ras gene family and human carcinogenesis. *Mutat. Res.* **195,** 255–271 (1988).
17. Finney, R. E. & Bishop, J. M. Predisposition to neoplastic transformation caused by gene replacement of H-ras1. *Science* **260,** 1524–1527 (1993).
18. Bremner, R. & Balmain, A. Genetic changes in skin tumor progression: correlation between presence of a mutant ras gene and loss of heterozygosity on mouse chromosome 7. *Cell* **61,** 407–417 (1990).
19. Yokota, J., Tsunetsugu-Yokota, Y., Battifora, H., Le Fevre, C. & Cline, M. J. Alterations of myc, myb, and rasHa proto-oncogenes in cancers are frequent and show clinical correlation. *Science* **231,** 261–265 (1986).
20. Bonner, R. F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* **278,** 1481–1483 (1997).
21. Salgia, R. & Skarin, A. T. Molecular abnormalities in lung cancer. *J. Clin. Oncol.* **16,** 1207–1217 (1998).
22. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88,** 593–602 (1997).

© 2001 Macmillan Magazines Ltd

23. Tanaka, N. *et al.* Cellular commitment to oncogene-induced transformation or apoptosis is dependent on the transcription factor IRF-1. *Cell* **77,** 829–839 (1994).
24. Jacks, T. *et al.* Tumor spectrum analysis in p53-mutant mice. *Curr. Biol.* **4,** 1–7 (1994).
25. Tuveson, D. A. & Jacks, T. Modeling human lung cancer in mice: similarities and shortcomings. *Oncogene* **18,** 5318–5324 (1999).
26. Yamashita, N., Minamoto, T., Ochiai, A., Onda, M. & Esumi, H. Frequent and characteristic K-ras activation in aberrant crypt foci of colon. Is there preference among K-ras mutants for malignant progression? *Cancer* **75,** 1527–1533 (1995).
27. Moser, A. R., Pitot, H. C. & Dove, W. F. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. *Science* **247,** 322–324 (1990).
28. Gibbs, J. B., Oliff, A. & Kohl, N. E. Farnesyltransferase inhibitors: Ras research yields a potential cancer therapeutic. *Cell* **77,** 175–178 (1994).
29. Sebti, S. & Hamilton, A. D. Inhibitors of prenyl transferases. *Curr. Opin. Oncol.* **9,** 557–561 (1997).
30. Lerner, E. C., Hamilton, A. D. & Sebti, S. M. Inhibition of Ras prenylation: a signaling target for novel anti-cancer drug design. *Anticancer Drug Des.* **12,** 229–238 (1997).

Supplementary information is available on *Nature's* World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank D. Jones, J. Whitsett, A. Mukherjee and G. Singh for advice and reagents. We also thank all laboratory members that provided input and advice on this project, as well as the Division of Comparative Medicine at MIT for their advice and care for the mice. This work was supported in part by grants from NCI, the Searle Scholars Program, and the MIT Charles Reed Fund.T.J. is an Associate Investigator of HHMI; D.A.T. is an HHMI Physician Postdoctoral Research Fellow.

Correspondence and requests for materials should be addressed to T.J. (e-mail: tjacks@mit.edu).

CONSTANS mediates between the circadian clock and the control of flowering in *Arabidopsis*

Paula Suárez-López*, Kay Wheatley*, Frances Robson*†, Hitoshi Onouchi*†, Federico Valverde* & George Coupland*‡

* *John Innes Centre, Norwich Research Park, Colney Lane, Norwich NR4 7UH, UK*
‡ *Max-Planck-Institut für Züchtungsforschung, Carl-von-Linné-Weg 10, 50829 Köln, Germany*

Flowering is often triggered by exposing plants to appropriate day lengths. This response requires an endogenous timer called the circadian clock to measure the duration of the day or night[1]. This timer also controls daily rhythms in gene expression and behavioural patterns such as leaf movements. Several *Arabidopsis* mutations affect both circadian processes and flowering time[2–10]; but how the effect of these mutations on the circadian clock is related to their influence on flowering remains unknown. Here we show that expression of *CONSTANS* (*CO*), a gene that accelerates flowering in response to long days[11], is modulated by the circadian clock and day length. Expression of a CO target gene, called *FLOWERING LOCUS T* (*FT*), is restricted to a similar time of day as expression of *CO*. Three mutations that affect circadian rhythms and flowering time alter *CO* and *FT* expression in ways that are consistent with their effects on flowering. In addition, the late flowering phenotype of such mutants is corrected by overexpressing *CO*. Thus, CO acts between the circadian clock and the control of flowering, suggesting mechanisms by which day length regulates flowering time.

Arabidopsis genes that affect flowering time have been identified and placed in genetic pathways[12]. *CO*, *LATE ELONGATED HYPOCOTYL* (*LHY*), *GIGANTEA* (*GI*), *FT* and the blue-light receptor *CRYPTOCHROME2* (*CRY2*, also named *FHA*) were assigned to the long-day (LD) pathway, which promotes flowering in response to long photoperiods. The autonomous pathway acts independently of day length, and includes *FCA* and *LUMINIDEPENDENS*. The circadian clock is also involved in regulating the floral transition[1]. *LHY*, *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*), *GI*, *EARLY FLOWERING 3* (*ELF3*), *TIMING OF CAB EXPRESSION 1* (*TOC1*), *ZEITLUPE* (*ZTL*) and *FKF1* (for flavin-binding, kelch repeat, F box) influence circadian rhythms and flowering time[2–10], but how their effects on these two processes are related is unknown. We have addressed this connection by studying the expression of *CO*, which was proposed to act in the same flowering-time pathway as the circadian-clock-related genes *LHY* and *GI* (refs 6, 13). *CO* encodes a putative transcription factor that is required to promote flowering under LD but not under short-day (SD) conditions[11].

The abundance of *LHY* and *GI* messenger RNA cycles with a 24-h rhythm and is controlled by the circadian clock[2,5,6]. Therefore, we tested whether *CO* mRNA abundance shows similar oscillations. *CO* mRNA levels varied under LD conditions, showing a broad peak between 12 h and dawn (Fig. 1a, c). The highest levels of mRNA occurred at 16 h and dawn, with a reproducible reduction at 20 h (Fig. 1a, c). As *CO* mRNA was reported to occur at lower abundance under SD than LD[11,14], we analysed whether the daily cycle in *CO* expression differed between these conditions. Under SD, the peak of *CO* expression was narrower than under LD and occurred between 12 and 20 h (Fig. 1a, d). The main differences between LD and SD were at 20 h and dawn (Fig. 1a, f). The higher abundance of *CO* mRNA under LD was most pronounced at dawn (Fig. 1a, f).

To investigate whether the daily oscillations in *CO* mRNA are controlled by the circadian clock, we analysed plants entrained under LD and transferred to constant light (LL). Under these conditions, *CO* mRNA levels continued to oscillate with a period of 24 h (Fig. 1b, e), showing that *CO* is regulated by the circadian clock. As we could not detect CO protein in wild-type plants, we studied plants overexpressing CO fusion proteins. The abundance of green fluorescent protein (GFP) or GFP–CO fusion protein was examined in plants expressing these proteins from the strong 35S promoter. GFP was at least 600 times more abundant than GFP–CO (Fig. 1g), although the abundance of their mRNAs differed by only 2–3-fold (Fig. 1h). Therefore, GFP–CO is unstable or poorly translated. Such instability of the CO protein suggests that its abundance closely follows that of its mRNA.

The *elf3* mutation causes early flowering[15] and disrupts circadian regulation of gene expression under LL[2,4,6], whereas a gain-of-function *lhy* mutation and loss-of-function *gi* mutations delay flowering and alter circadian rhythms[2,5,6,13]. Therefore, we tested *CO* mRNA abundance in these mutants under LD and SD (Fig. 2). In the *gi-3* mutant, *CO* mRNA cycled in the same phase as in wild-type plants but at lower amplitude (Fig. 2a, d), consistent with the effect of *gi-3* on other mRNAs under light/dark cycles[2]. In the *lhy* mutant, *CO* mRNA abundance was reduced and its rhythm was altered, showing a narrow peak in expression at a different phase as compared with wild type (Fig. 2a, d). Another circadian clock regulated gene, *CCR2* (for cold, circadian rhythm, and RNA binding; refs 16, 17), also showed an altered rhythm in *lhy* mutants under LD. In wild-type and *co-2* mutants, the peak in *CCR2* expression occurred at 12 h, but in *lhy* mutants the peak was at 0 h (Fig. 2l, m). Therefore, *lhy* may have a general effect on the phase of expression of clock-regulated genes under LD.

In *lhy* and *gi-3* mutants entrained under LD and transferred to LL, *CO* mRNA abundance was decreased and appeared to be arrhythmic, although the presence of low-amplitude oscillations could not be excluded (Fig. 2e). These data indicate that *lhy* and *gi-3* affect the circadian regulation of *CO* and reduce *CO* mRNA abundance under LD. In contrast, *elf3-1* caused an increase in *CO* mRNA levels in LD at all times tested (Fig. 2b, f) and in SD at least during the light period (Fig. 2c, g). Thus, late flowering in *lhy* and *gi-3* correlates

† Present addresses: School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand (F.R.); Division of Applied Bioscience, Graduate School of Agriculture, Hokkaido University, Sapporo 060-8589, Japan (H.O.).

© 2001 Macmillan Magazines Ltd

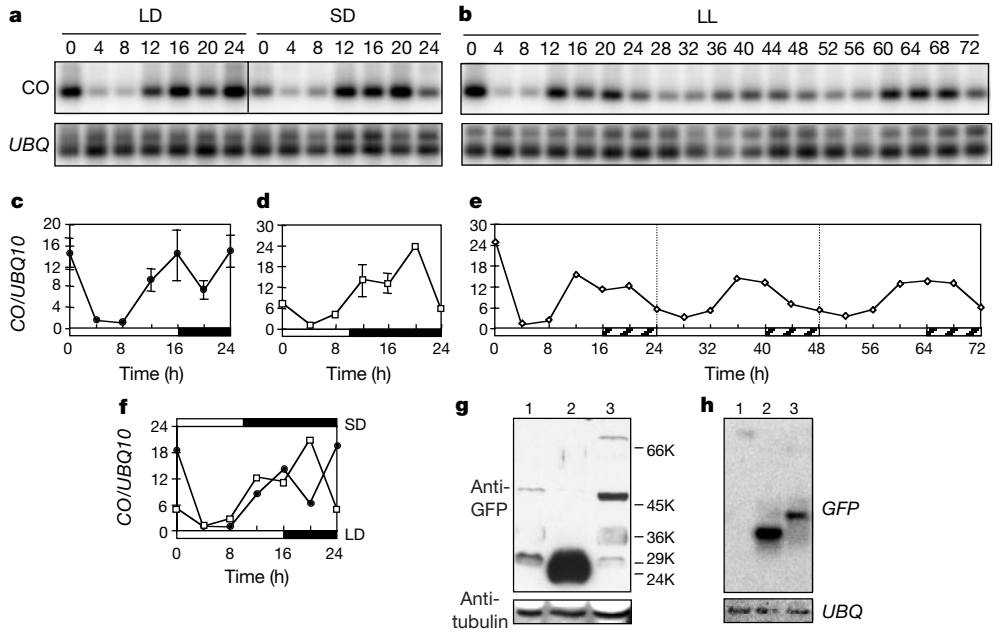


Figure 1 *CO* expression in wild type-plants. **a**, **b**, *CO* mRNA abundance in plants grown in LD and SD (**a**) or entrained in LD and transferred to LL (**b**). Samples were collected at the times shown after dawn (time 0). **c**–**f**, Quantification of *CO* mRNA from the experiments shown in **a** (left), **a** (right), **b** and **a**, respectively. **c**, Mean ± s.e.m. of four independent experiments in LD. **d**, Mean ± s.e.m. of three independent experiments in SD. **e**, Representative result of three independent experiments in LL. **f**, Representative result of two independent experiments. LD, filed circles; SD, open squares; LL, open diamonds. Open, filled and hatched bars represent light, dark and subjective dark periods, respectively. **g**, Western blot analyses of WT (1), *35S::GFP* (2) and *35S::GFP-CO* (3) seedling extracts using anti-GFP and anti-tubulin antibodies. One hundred micrograms of protein extract was used, except for *35S::GFP* in the GFP blot (10 μg). The expected sizes of GFP and GFP-CO proteins are approximately 26 and 72 kDa, respectively. **h**, Northern blot analyses of WT (1), *35S::GFP* (2) and *35S::GFP–CO* (3) seedling RNA using *GFP* and *UBQ10* probes.

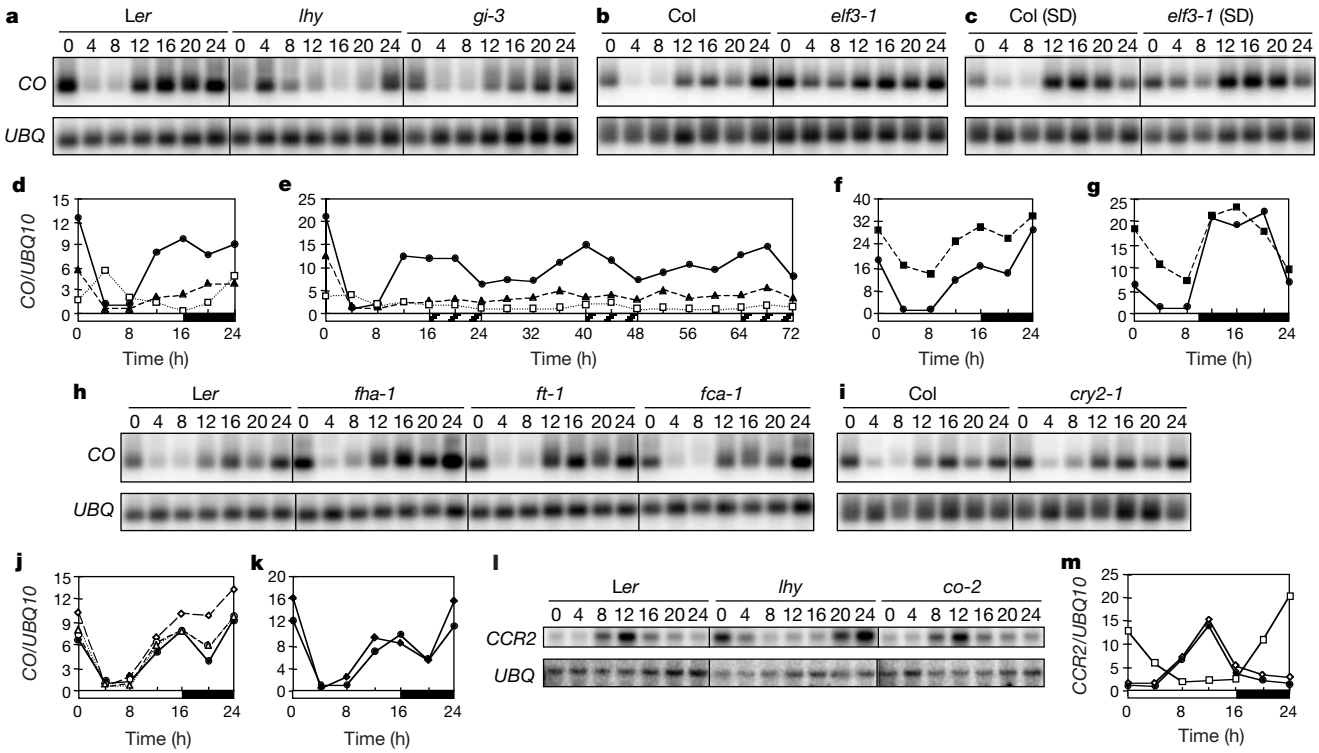


Figure 2 *CO* expression in flowering-time mutants. **a**–**c**, **h**, **i**, Analysis of *CO* mRNA in plants grown in LD (**a**, **b**, **h**, **i**) and SD (**c**). Mutants and wild types (L*er* or Columbia, Col) are indicated. Samples were collected at the times shown after dawn. **d**, **f**, **g**, **j**, **k**, Quantification of the data shown in **a**–**c**, **h** and **i**, respectively. WT, filled circles; *lhy*, open squares; *gi-3*, filled triangles; *elf3-1*, filled squares; *fha-1*, open diamonds; *ft-1*, open circles; *fca-1*, open triangles; *cry2-1*, filled diamonds. **e**, Quantification of *CO* mRNA abundance in WT (filled circles), *lhy* (open squares) and *gi-3* (filled triangles) plants entrained in LD and transferred to LL. **l**, Northern blot analysis of *CCR2* mRNA in plants grown in LD. **m**, Quantification of *CCR2* mRNA abundance from the blots shown in **l**. WT, filled circles; *lhy*, open squares; *co-2*, open diamonds. Open, filled and hatched bars represent light, dark and subjective dark periods, respectively. Panels **d**–**g** are representative of two independent experiments.

© 2001 Macmillan Magazines Ltd

with reduced *CO* levels, whereas early flowering in *elf3-1* correlates with elevated *CO* expression. *LHY*, *GI* and *ELF3* may therefore regulate flowering by modulating transcription of *CO*.

To establish whether late flowering of *lhy* and *gi-3* was caused by reduction in *CO* mRNA, *CO* was expressed from the 35S promoter in these mutant backgrounds. *35S::CO* transgenic plants flower early in LD and SD[18]. Both *35S::CO lhy* and *35S::CO gi-3* plants flowered much earlier than *lhy* and *gi-3* and as early as *35S::CO* (Table 1). This supports the proposal that less *CO* mRNA in *lhy* and *gi* mutants (Fig. 2a, d) causes their late-flowering phenotype. We did not test genetically whether increased *CO* expression in *elf3* causes early flowering because *co* and *elf3* mutations are available only in different *Arabidopsis* ecotypes; however, this is supported by the suppression of the early flowering of *elf3* by *gi* (ref. 19).

In the *ft-1* and *fca-1* late-flowering mutants, *CO* mRNA abundance was not affected in LD (Fig. 2h, j), in agreement with *FT* and *FCA* acting downstream of *CO* and in different pathways, respectively[13,20,21]. Previously, *CO* mRNA levels were reported to be increased by CRY2 (ref. 14), but under LD we detected no reduction of *CO* mRNA levels in the *fha-1* mutant, which is an allele of *cry2* (Fig. 2h, j). We further tested the relationship between *CO* and *CRY2* using *cry2-1* and *fha-1* alleles, by growing plants in LD, SD as well as true LD (see Methods) and by extracting RNA at two developmental times. In all cases, however, *CO* mRNA abundance was similar in the mutants and wild type (Fig. 2i, k; and data not shown). Nevertheless, *35S::CO fha-1* plants flowered at the same time as *35S::CO* (Table 1), indicating that overexpression of *CO* corrects the small effect of *fha-1* on flowering. Because under our conditions *FHA* probably does not regulate *CO* transcription, *35S::CO* may correct *fha-1* by a mechanism independent of *CO* transcriptional regulation.

CO has been shown to promote flowering by activating the expression of *FT* and *SUPPRESSOR OF OVEREXPRESSION OF CO 1* (*SOC1*; refs 20–22). Mutations in *FT* delay flowering in wild-type and *35S::CO* plants[13,18]. Therefore, we tested whether *FT* mRNA showed a similar rhythm to *CO* mRNA. In wild-type plants grown under LD, *FT* mRNA peaks at 20 h (Fig. 3a, c), when *CO* mRNA abundance is also high (Fig. 1). This peak is absent in *co-2*, *gi-3* and *lhy* mutants (Fig. 3a, c), which is consistent with the activation of *FT* by CO (refs. 20–22), and reduced *CO* expression in *lhy* and *gi* mutants (Fig. 2a, d). In addition, in the *lhy* mutant the narrow peak in the *CO* mRNA level at 4 h (Fig. 2a, d) correlates with a slightly later low amplitude peak in *FT* mRNA (Fig. 3a, c). Abundance of *FT* mRNA is increased in *35S::CO*, *35S::CO lhy* and *elf3-1* plants (Fig. 3a, b, d, e), consistent with their early flowering phenotype (Table 1 and ref. 15).

Table 1 Effect of *35S::CO* on flowering time in different mutant backgrounds

		Rosette leaves	Cauline leaves	Total leaf number	Days to flowering
WT (*Ler*)	LD	5.5 ± 0.1	3.0 ± 0.1	8.5 ± 0.2	17.0 ± 0.2
	SD	28.2 ± 0.4	10.7 ± 0.1	38.8 ± 0.4	43.3 ± 0.6
35S::CO	LD	2.5 ± 0.2	2.0 ± 0.1	4.5 ± 0.2	13.5 ± 0.3
	SD	2.4 ± 0.1	2.0 ± 0.1	4.5 ± 0.1	15.1 ± 0.3
35S::CO lhy	LD	2.6 ± 0.2	1.7 ± 0.2	4.3 ± 0.2	ND*
	SD	3.0 ± 0.1	1.9 ± 0.1	5.0 ± 0.2	ND
lhy	LD	17.3 ± 1.5	6.8 ± 0.5	24.1 ± 1.7	ND
	SD	22.9 ± 2.6†	7.4 ± 0.6†	30.3 ± 2.6†	ND
35S::CO gi-3	LD	2.3 ± 0.2	2.0 ± 0.0	4.3 ± 0.2	12.9 ± 0.4
	SD	2.8 ± 0.1	1.5 ± 0.1	4.3 ± 0.1	14.6 ± 0.5
gi-3	LD	15.6 ± 0.4‡	8.4 ± 0.8‡	24.0 ± 1.0‡	39.2 ± 1.5‡
	SD	ND	ND	26.0 ± 0.6	64.5 ± 1.2
35S::CO fha-1	LD	2.6 ± 0.1	2.0 ± 0.1	4.5 ± 0.1	13.5 ± 0.3
	SD	2.7 ± 0.1	1.6 ± 0.1	4.3 ± 0.1	14.8 ± 0.3
fha-1	LD	6.5 ± 0.1	3.8 ± 0.1	10.3 ± 0.2	19.6 ± 0.2
	SD	24.9 ± 0.4	10.5 ± 0.4	35.3 ± 0.8	41.8 ± 0.6

In each case, at least 10 plants were analysed, except where indicated.
* ND, not determined.
† Seven plants were analysed.
‡ Five plants were analysed.

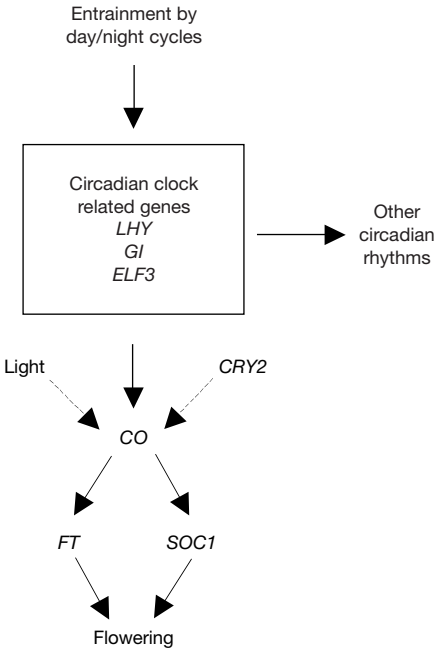


Figure 4 Model of the long-day flowering time pathway. *LHY*, *GI* and *ELF3* influence circadian rhythms and flowering. Mutations in these genes alter the rhythms in *CO* expression. No effect of CRY2 on *CO* mRNA was detected, but CRY2 may regulate *CO* at the post-transcriptional level (dotted line) or act independently. Regulation of CO by light under long photoperiods is proposed as a mechanism by which CO activity is restricted to long days (dotted line). CO promotes flowering through the activation of *FT* and *SOC1* (ref. 21), and *FT* mRNA abundance cycles in a similar phase to *CO* mRNA.

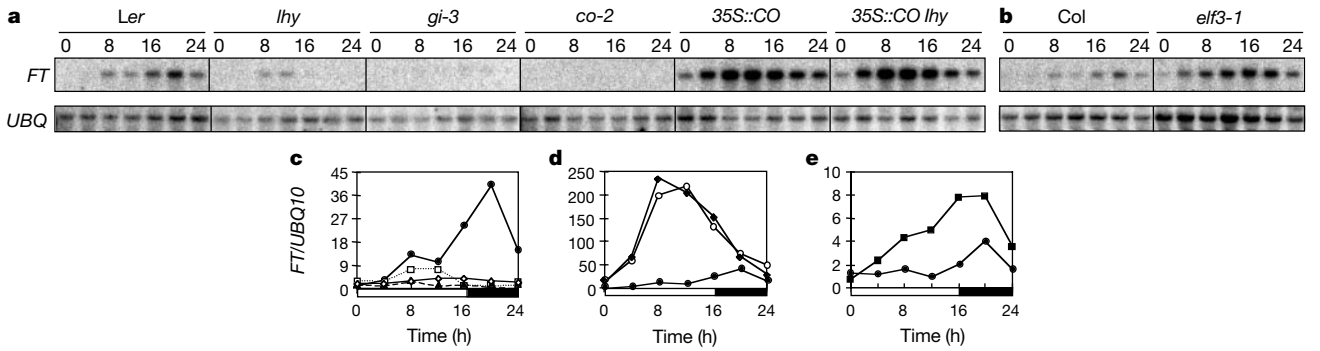


Figure 3 Analysis of *FT* mRNA levels. **a**, **b**, Northern blot analysis of *FT* mRNA abundance in plants grown in LD. **c**–**e**, Quantification of *FT* mRNA abundance from the blots shown in **a** and **b**. WT, filled circles; *lhy*, open squares; *gi-3*, filled triangles; *co-2*, open diamonds; *35S::CO*, open circles; *35S::CO lhy*, filled diamonds; *elf3-1*, filled squares. Open and filled bars represent light and dark periods, respectively.

© 2001 Macmillan Magazines Ltd

LHY, *GI* and *CO* mRNA abundance oscillates in a circadian rhythm[2,5,6] (Fig. 1). Unlike LHY and GI, however, CO does not appear to be involved in circadian clock function[23] (Fig. 2l, m), although it does show homology with TOC1, a protein implicated in circadian oscillator function[10]. We propose that CO mediates between the circadian oscillator and activation of the flowering-time gene *FT* (Fig. 4). In wild-type plants, CO promotes flowering under LD but not SD. Under LD, *CO* mRNA abundance is high at the end and the beginning of the photoperiod, whereas under SD the peak in *CO* mRNA occurs only in darkness (Fig. 1). Thus, if the translation, activity or stability of the CO protein is regulated by light, this might provide a mechanism by which *CO* promotes flowering specifically under LD. Although our data do not exclude the possibility of post-transcriptional regulation of CO in LD independently of light, a mechanism involving light regulation is supported by the observation that in *35S::CO* plants growing in LD the peak in *FT* mRNA abundance is much higher in light than dark (Fig. 3a, d), despite the 35S promoter being active at all times. Similarly, the peak in *FT* mRNA at 20 h in LD might be caused by the exposure of CO protein to light during the preceding photoperiod. Indeed, the abundance of *FT* mRNA rises steeply around 16 h when *CO* expression is high. Such temporal control may be important in regulating FT function. Together, our data indicate that circadian clock regulation of *CO* may represent a light-sensitive circadian rhythm proposed to underlie the photoperiodic control of flowering[1].

Although circadian output genes have been identified in several systems (for example, see refs 24 and 25), their functions are mostly unknown. Recently, the *Drosophila takeout* gene was proposed to link the circadian clock with feeding behaviour[26]. *CO* may have a similar role in an output pathway that integrates day-length perception and timekeeping mechanisms to promote flowering. Examining how the daily rhythms in *CO* mRNA levels are generated and the possible role of light in post-transcriptional regulation of CO will further elucidate the mechanism by which plants respond to day length. □

Methods

Plant material and growth conditions

We used the Landsberg *erecta* (*Ler*) ecotype of *Arabidopsis thaliana* unless otherwise indicated. The *gi-3*, *fha-1*, *ft-1* and *fca-1* mutants[13] were provided by M. Koornneef. The *lhy* and *35S::CO* plants have been described[6,18]. The *elf3-1* (ref. 15) and *cry2-1* (ref. 14) mutants, both in Columbia ecotype, were gifts from R. Meeks-Wagner and C. Lin, respectively. We obtained *35S::CO lhy*, *35S::CO gi* and *35S::CO fha-1* lines by crossing either *35S::CO* or *35S::CO co-2* plants to the corresponding mutants. The *35S::CO lhy* line used also carries *co-2*, which does not affect the flowering time of *35S::CO* plants. The *35S::GFP* plants were provided by R. Sablowski. The *35S::GFP-CO* plants will be described elsewhere.

Plants were grown in soil in controlled environment rooms under LD (10-h light/6-h day extension/8-h dark) or SD (10-h light/14-h dark) as described[11], or under true LD (16-h light (400-W metal halide power star lamps supplemented with 100-W tungsten halide lamps)/ 8-h dark). For LL experiments, plants were grown on agar plates under true LD (16-h light/8-h dark) for 8 d and then transferred to LL at dawn[6].

Analysis of *CO* mRNA abundance

Sample collection started at dawn of day 8 for all experiments. We used the aerial parts of seedlings grown in soil or whole seedlings grown on plates. Tissue was ground in liquid nitrogen, resuspended in extraction buffer (7.5 M guanidine hydrochloride, 25 mM sodium citrate, 5 mM dithiothreitol, pH 7.0) and extracted with phenol:chloroform:isoamyl alcohol. RNA was precipitated with acetic acid as described[27], resuspended in water and treated with DNase I (Pharmacia) according to the manufacturer's instructions. Because the *CO* transcript is very rare, we carried out RNA analysis by polymerase chain reaction with reverse transcription (RT–PCR). For synthesis of complementary DNA, 5 μg of total RNA was primed using the dT_{17} primer as described[28]. cDNAs were diluted to 200 μl with water, and 5 μl of diluted cDNA was used for PCR amplification. A *CO* fragment was amplified using primers CO53, 5′-ACGCCATCAGCGAGTTCC-3′, and COoli9, 5′-AAATGTATGCGTTATGGT-TAATGG-3′. A *UBQ10* fragment was amplified[29] and used as a control to normalize the amounts of cDNA. Several numbers of cycles were used for both *CO* and *UBQ10* PCR to determine the exponential range of amplification. Then, 25 and 20 cycles were used for *CO* and *UBQ10* PCR, respectively, in all the experiments shown. We separated PCR products on an agarose gel, transferred them to a Hybond NX nylon membrane (Amersham) and hybridized them with radioactively labelled probes according to standard procedures. Full-length *CO* and *UBQ10* cDNAs were used as probes. Images were visualized using a PhosphorImager (Molecular Dynamics), and band intensities were quantified using ImageQuant software (Molecular Dynamics). Values were represented relative to the lowest value of the wild-type samples after normalization to the *UBQ10* control.

Analysis of gene expression

RNA (10 μg) was separated on 1.2% agarose denaturing formaldehyde gels and transferred to Hybond NX nylon membranes. Hybridization with radioactively labelled probes was done in 0.3 M sodium phosphate buffer, pH 7.0, 7% SDS, 1 mM EDTA, 1% bovine serum albumin overnight at 65 °C. The blot was washed for 20 min at 65 °C with 2× SSC, 0.1% SDS and twice for 10 min at 65 °C with 0.2× SSC, 0.1% SDS. We used full-length *GFP* and *CCR2* cDNAs as probes. The *UBQ10*- and *FT*-specific probes have been described[21,30]. Images were visualized, intensities quantified and values represented as described above.

Immunoblot analysis

Protein was prepared using EZ buffer[31] from 7-day-old seedlings grown in LL. We loaded 100 μg of extract per lane on discontinuous SDS-acrylamide gels, and performed western blots using commercially available GFP antibodies (Santa Cruz Biotechnology).

Measurement of flowering time

Flowering time was measured by scoring the number of rosette and cauline leaves on the main stem. The number of days from sowing until flower buds were visible by eye at the centre of the rosette was also recorded. Data are expressed as means ± s.e.m.

Received 2 December 2000; accepted 2 February 2001.

1. Thomas, B. & Vince-Prue, D. *Photoperiodism in Plants* (Academic, London, 1997).
2. Fowler, S. *et al. GIGANTEA*: a circadian clock-controlled gene that regulates photoperiodic flowering in *Arabidopsis* and encodes a protein with several possible membrane-spanning domains. *EMBO J.* **18,** 4679–4688 (1999).
3. Green, R. M. & Tobin, E. M. Loss of the circadian clock-associated protein 1 in *Arabidopsis* results in altered clock-regulated gene expression. *Proc. Natl Acad. Sci. USA* **96,** 4176–4179 (1999).
4. Hicks, K. A. *et al.* Conditional circadian dysfunction of the *Arabidopsis early-flowering-3* mutant. *Science* **274,** 790–792 (1996).
5. Park, D. H. *et al.* Control of circadian rhythms and photoperiodic flowering by the *Arabidopsis GIGANTEA* gene. *Science* **285,** 1579–1582 (1999).
6. Schaffer, R. *et al.* The *late elongated hypocotyl* mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering. *Cell* **93,** 1219–1229 (1998).
7. Wang, Z. -Y. & Tobin, E. M. Constitutive expression of the *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) gene disrupts circadian rhythms and suppresses its own expression. *Cell* **93,** 1207–1217 (1998).
8. Nelson, D. C., Lasswell, J., Rogg, L. E., Cohen, M. A. & Bartel, B. *FKF1*, a clock-controlled gene that regulates the transition to flowering in *Arabidopsis*. *Cell* **101,** 331–340 (2000).
9. Somers, D. E., Schultz, T. F., Milnamow, M. & Kay, S. A. *ZEITLUPE* encodes a novel clock-associated PAS protein from *Arabidopsis*. *Cell* **101,** 319–329 (2000).
10. Strayer, C. *et al.* Cloning of the *Arabidopsis* clock gene *TOC1*, an autoregulatory response regulator homolog. *Science* **289,** 768–771 (2000).
11. Putterill, J., Robson, F., Lee, K., Simon, R. & Coupland, G. The *CONSTANS* gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell* **80,** 847–857 (1995).
12. Simpson, G. G., Gendall, A. R. & Dean, C. When to switch to flowering. *Annu. Rev. Cell Dev. Biol.* **15,** 519–550 (1999).
13. Koornneef, M., Hanhart, C. J. & van der Veen, J. H. A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. *Mol. Gen. Genet.* **229,** 57–66 (1991).
14. Guo, H., Yang, H., Mockler, T. C. & Lin, C. Regulation of flowering time by *Arabidopsis* photoreceptors. *Science* **279,** 1360–1363 (1998).
15. Zagotta, M. T. *et al.* The *Arabidopsis ELF3* gene regulates vegetative photomorphogenesis and the photoperiodic induction of flowering. *Plant J.* **10,** 691–702 (1996).
16. Carpenter, C. D., Kreps, J. A. & Simon, A. E. Genes encoding glycine-rich *Arabidopsis thaliana* proteins with RNA- binding motifs are influenced by cold treatment and an endogenous circadian rhythm. *Plant Physiol.* **104,** 1015–1025 (1994).
17. Heintzen, C., Nater, M., Apel, K. & Staiger, D. *At*GRP7, a nuclear RNA-binding protein as a component of a circadian- regulated negative feedback loop in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **94,** 8515–8520 (1997).
18. Onouchi, H., Igeño, M. I., Périlleux, C., Graves, K. & Coupland, G. Mutagenesis of plants over-expressing *CONSTANS* demonstrates novel interactions among Arabidopsis flowering-time genes. *Plant Cell* **12,** 885–900 (2000).
19. Chou, M. -L. & Yang, C. -H. Late-flowering genes interact with early-flowering genes to regulate flowering time in *Arabidopsis thaliana*. *Plant Cell Physiol.* **40,** 702–708 (1999).
20. Kardailsky, I. *et al.* Activation tagging of the floral inducer *FT*. *Science* **286,** 1962–1965 (1999).
21. Samach, A. *et al.* Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. *Science* **288,** 1613–1616 (2000).
22. Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M. & Araki, T. A pair of related genes with antagonistic roles in mediating flowering signals. *Science* **286,** 1960–1962 (1999).
23. Somers, D. E. & Kay, S. A. in *Biological Rhythms and Photoperiodism in Plants* (eds Lumsden, P. J. & Millar, A. J.) 81–98 (BIOS Scientific, Oxford, 1998).
24. Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96,** 271–290 (1999).
25. Wager-Smith, K. & Kay, S. A. Circadian rhythm genetics: from flies to mice to humans. *Nature Genet.* **26,** 23–27 (2000).
26. Sarov-Blat, L., So, W. V., Liu, L. & Rosbash, M. The *Drosophila takeout* gene is a novel molecular link between circadian rhythms and feeding behavior. *Cell* **101,** 647–656 (2000).
27. Logemann, J., Schell, J. & Willmitzer, L. Improved method for the isolation of RNA from plant tissues. *Anal. Biochem.* **163,** 16–20 (1987).

© 2001 Macmillan Magazines Ltd

28. Frohman, M. A., Dush, M. K. & Martin, G. R. Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA* **85,** 8998–9002 (1988).
29. Blázquez, M. A. & Weigel, D. Independent regulation of flowering by phytochrome B and gibberellins in *Arabidopsis*. *Plant Physiol.* **120,** 1025–1032 (1999).
30. Wang, Z. -Y. *et al.* A Myb-related transcription factor is involved in the phytochrome regulation of an *Arabidopsis Lhcb* gene. *Plant Cell* **9,** 491–507 (1997).
31. Martínez-García, J. F., Monte, E. & Quail, P. H. A simple, rapid and quantitative method for preparing *Arabidopsis* protein extracts for immunoblot analysis. *Plant J.* **20,** 251–257 (1999).

Acknowledgements

We thank L. Wright for RNA samples; M. I. Igeño for the cross *gi-3* × *35S::CO*; R. W. M. Sablowski for the *35S::GFP* plants; M. M. R. Costa and M. Piñeiro for the *35S::GFP-CO* plants; M. Koornneef, C. Lin and R. Meeks-Wagner for mutant seeds; C. Andronis for the *UBQ10* cDNA probe; A. Samach for scientific discussions; and C. Dean, T. Mizoguchi, P. H. Reeves, A. Samach and G. G. Simpson for comments on the manuscript. We are grateful to the laboratories of J. Putterill and E. M. Tobin for sharing unpublished results, and to C. Andronis and M. A. Blázquez for advice with the *UBQ10* controls. P.S.-L. was supported by a fellowship from the Human Frontier Science Program Organization and a Marie Curie fellowship, and F.V. by a FEBS long-term fellowship.

Correspondence and requests for materials should be addressed to G.C. (e-mail: george.coupland@bbsrc.ac.uk or coupland@mpiz-koeln.mpg.de).

Structure of the gating domain of a Ca^{2+}-activated K^+ channel complexed with Ca^{2+}/calmodulin

Maria A. Schumacher*, Andre F. Rivard*, Hans Peter Bächinger†‡ & John P. Adelman*

** Vollum Institute, Oregon Health Sciences University, Portland, Oregon 97201-3098, USA*
† Shriners Hospital for Children, Research Unit, 3101 S W Sam Jackson Park Road, Portland, Oregon 97201, USA
‡ Department of Biochemistry and Molecular Biology, Oregon Health Sciences University, Portland, Oregon 97201, USA

Small-conductance Ca^{2+}-activated K^+ channels (SK channels)[1,2] are independent of voltage and gated solely by intracellular Ca^{2+}. These membrane channels are heteromeric complexes that comprise pore-forming α-subunits and the Ca^{2+}-binding protein calmodulin (CaM). CaM binds to the SK channel through the CaM-binding domain (CaMBD), which is located in an intracellular region of the α-subunit immediately carboxy-terminal to the pore[3,4]. Channel opening is triggered when Ca^{2+} binds the EF hands in the N-lobe of CaM[4]. Here we report the 1.60 Å crystal structure of the SK channel CaMBD/Ca^{2+}/CaM complex. The CaMBD forms an elongated dimer with a CaM molecule bound at each end; each CaM wraps around three α-helices, two from one CaMBD subunit and one from the other. As only the CaM N-lobe has bound Ca^{2+}, the structure provides a view of both calcium-dependent and -independent CaM/protein interactions. Together with biochemical data, the structure suggests a possible gating mechanism for the SK channel.

The structure of the CaMBD from rat SK2 (residues 395–490) complexed to Ca^{2+}/CaM was determined by multiple isomorphous replacement (MIR) and includes CaM residues 1–147, the non-helical CaM linker region, two Ca^{2+} ions (bound in the CaM N-lobe) and CaMBD residues 413–489 (Fig. 1a, b). No density was observed for CaMBD residues 395–412, which connects to the sixth transmembrane helix (S6) of the channel. In the complex, the CaMBD consists of two long α-helices, α1 (residue 413–440) and α2 (residues 446–489), connected by a loop (residues 441–445); CaM contains two EF-hand-containing lobes, the N-lobe and the C-lobe, which are connected by a linker region (residues 75–80). The crystal packing shows a clear CaMBD dimer, created by the side-by-side antiparallel interaction of α2 and α2′ (where the prime indicates the other subunit), which buries ~400 Å^2 of each monomer surface. No other CaMBD/CaMBD contacts are observed in the crystal. Two CaM molecules are bound to the CaMBD dimer such that each CaM molecule grips an end of the dimer (Fig. 1a). In this interaction, each molecule of CaM contacts both subunits of the CaMBD dimer, forming a highly elongated complex with dimensions of about 80 × 54 × 50 Å. On binding, CaM nearly engulfs the CaMBD dimer, burying over 80% of its surface area.

In other structures of Ca^{2+}/CaM/peptide complexes, CaM binds a single peptide α-helix, encasing the helix between its two lobes[5–9]. However, the CaMBD contains no identifiable CaM interaction motif, either Ca^{2+} dependent or independent[10], in its CaM-binding sequence. Indeed, the CaMBD/Ca^{2+}/CaM complex is distinct from any previously described CaM/peptide complex because CaM binds three α-helices instead of one, and the N-lobe and C-lobe of each CaM molecule contact different CaMBD monomers. Furthermore, as Ca^{2+} is not bound in the C-lobe EF hands, the CaMBD/Ca^{2+}/CaM structure details both Ca^{2+}-dependent and Ca^{2+}-independent CaM interactions in a single complex. This finding is consistent with biochemical data[4]. Specifically, in contrast to findings that suggested the importance of Ca^{2+} binding to the CaM C-lobe in SK activation[3], our subsequent, more extensive studies[4] showed that C-lobe Ca^{2+}-coordinating residues can be mutated and Ca^{2+} binding destroyed without any effect on SK gating. Ca^{2+}-coordinating residues in the

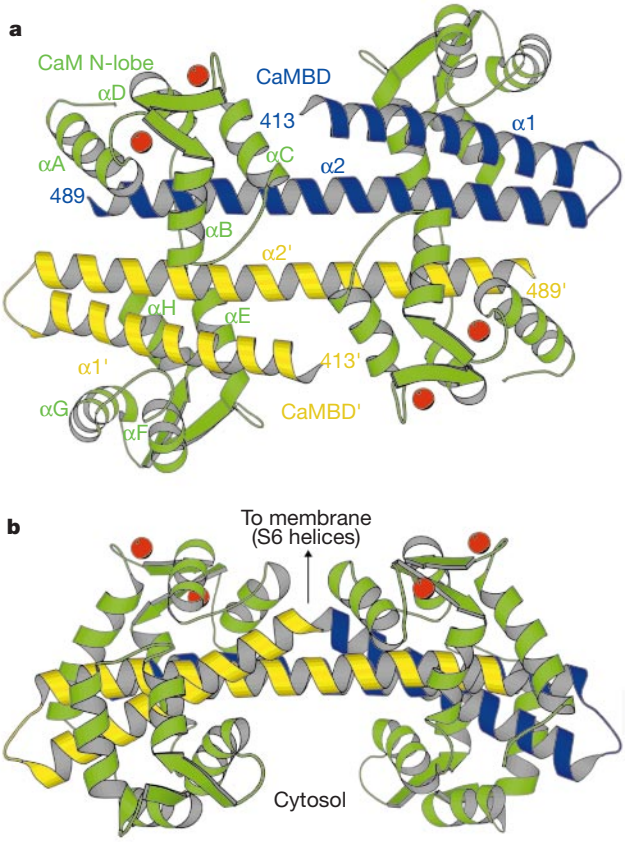


Figure 1 Structure of the CaMBD/Ca^{2+}/CaM complex. **a**, Ribbon diagram of the CaMBD/Ca^{2+}/CaM dimeric complex. CaMBD subunits are in blue and yellow, CaM molecules are in green, and the Ca^{2+} ions are in red. Secondary structural elements, the CaM linker and the first and last observed residues in the CaMBD are labelled. **b**, View in **a** rotated by 90° showing the orientation of the complex relative to the membrane. Arrow indicates the positions of the first observed residue of each of the CaMBD monomers that are linked to the S6 pore helices. Figure generated with MOLSCRIPT[29].

© 2001 Macmillan Magazines Ltd

N-lobe were, however, shown to be essential for gating[4].

The CaMBD/Ca^{2+}/CaM complex is stabilized by an extensive number of contacts (Fig. 2a–c; and Supplementary Information Fig. 1). In addition to overall electrostatic complementarity between the positive CaMBD (pI 10.5; charge +14.7) and acidic CaM surfaces, CaM is anchored onto the CaMBD by several hydrogen-bond tethers (Fig. 2a–c; and Supplementary Information Fig. 1). Although these interactions anchor the CaM onto the CaMBD, the specific binding interface between the two proteins is primarily hydrophobic. In the CaM N-lobe, residues Phe 12, Phe 19, Val 35, Met 36, Leu 39, Phe 68, Met 71 and Met 72 create a complementary hydrophobic patch for interaction with CaMBD residues Leu 480, Leu 483 and Ala 484 (Fig. 2a). The N-lobe/CaMBD interaction is similar to other Ca^{2+}/CaM/peptide structures[5–8] in that it contains a 'hydrophobic anchor point', namely, the side chain of Leu 480, which reaches into the N-lobe core (Fig. 2a). However, a unique feature of the N-lobe/CaMBD interaction is that each N-lobe interacts with two CaMBD subunits; in addition to the main N-lobe/CaMBD interaction involving the Leu 480 prong, a short CaM N-lobe loop (Gly 40 to Pro 43) reaches around and contacts the other subunit of the CaMBD dimer (Fig. 2a, b; and Supplementary Information Fig. 1). Thus, the CaMBD/N-lobe interaction stabilizes the dimeric complex.

This N-lobe loop along with a second C-lobe loop comprising Asn 111 to Leu 116 encase the 'front' of the CaMBD dimer (Figs 1a and 2b). In this second loop, the side chain of Glu 114 has an important role in anchoring the CaM C-lobe onto the CaMBD (Fig. 2b; and Supplementary Information Fig. 1). In other Ca^{2+}/CaM/peptide structures, these two loops are brought into close proximity on peptide binding and interact with each other and not the peptide. In contrast, CaM complexed to the CaMBD adopts a highly elongated conformation by the extension of its linker region, which is essential in enabling a single CaM to wrap around three α-helices.

The CaMBD/C-lobe complex represents a new type of CaM interaction. This results from the distinct conformation adopted by the Ca^{2+}-free C-lobe as compared with the Ca^{2+}-bound state[11,12]. Superimposition of Cα residues 81–146 of the CaMBD-complexed CaM C-lobe onto those of the Ca^{2+}/CaM C-lobe[12] results in an r.m.s. deviation of 2.80 Å, compared with an r.m.s. deviation of 0.58 Å obtained from a similar overlay of Cα residues 10–67 of the corresponding N-lobes (Fig. 3a). Thus the calcified N-lobe in the CaMBD/Ca^{2+}/CaM complex takes the Ca^{2+}-bound conformation, whereas the C-lobe is structurally distinct, perhaps retaining the Ca^{2+}-free or apo structure (Fig. 3a, b).

Indeed, comparing the structure of the C-lobe from the CaMBD/Ca^{2+}/CaM complex to C-lobe structures of apoCaM solved by NMR[13–15] reveals the same overall trends: a re-orientation of helical pairs αG/αF with respect to αE/αH and the subsequent repacking of the hydrophobic core, which leads to the creation of a more compact lobe as compared with the Ca^{2+}-bound condition. Furthermore, superimposing C-lobe Cα residues 81–146 of the CaMBD-bound C-lobe onto those of the NMR apoCaM C-lobe structure[15]

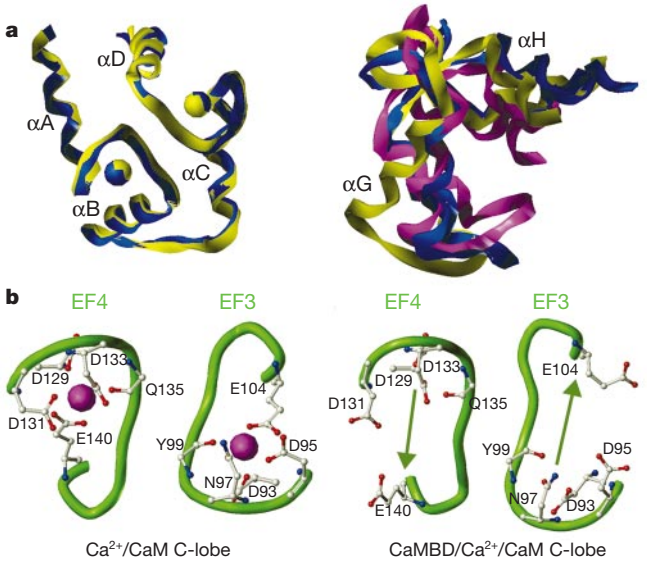


Figure 3 Structure of CaM lobes in the CaMBD/Ca^{2+}/CaM complex. **a**, Left, Cα superimposition of N-lobe residues 10–67 of the CaMBD/Ca^{2+}/CaM complex (blue) onto those of the Ca^{2+}/CaM complex[12] (yellow). Ca^{2+} ions are shown as spheres. Right, Cα superimposition of C-lobe residues 81–146 of the CaMBD/Ca^{2+}/CaM complex (blue) and the NMR apoCaM C-lobe structure[15] (magenta) onto that of the Ca^{2+}/CaM complex[12] (yellow). **b**, Comparison of the C-lobe EF hand regions of the Ca^{2+}/CaM[12] and the CaMBD/Ca^{2+}/CaM complexes. Ca^{2+}-coordinating residues are shown as sticks. Note the displacements of Ca^{2+}-coordinating residues Asp 95, Asp 131, Glu 104 and Glu 140 out of the expanded Ca^{2+}-coordination sphere in the CaMBD complex.

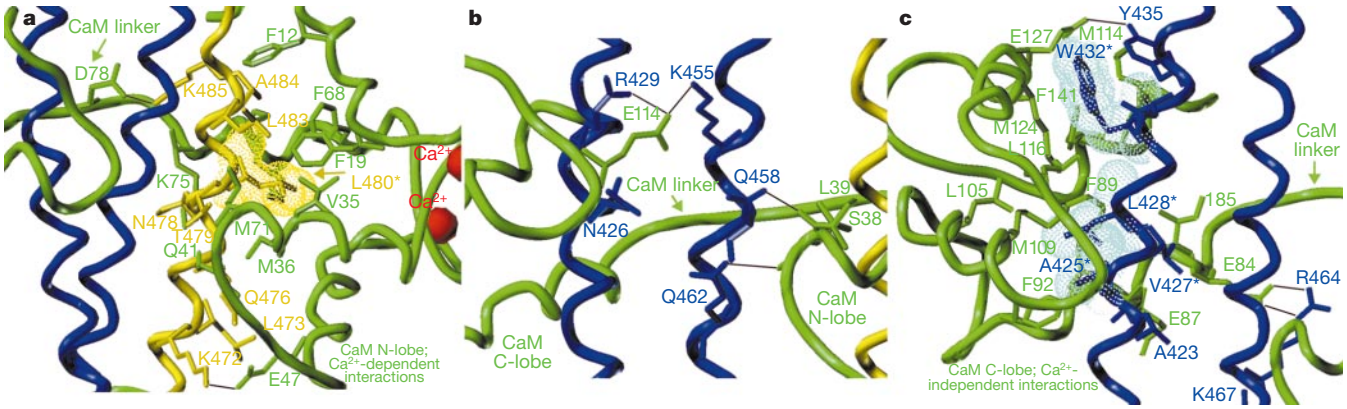


Figure 2 Interactions between CaMBD and CaM. **a**, Ca^{2+}-dependent interactions between the CaMBD and the CaM N-lobe. CaMBD subunits and CaM are coloured as in Fig. 1. Hydrogen-bond interactions are indicated as black lines. The van der Waals surface (yellow) is shown for the single 'leucine prong'. **b**, Tethering contacts between CaMBD and CaM highlight interactions between the CaM N-lobe and the second CaMBD dimer (blue), where the contacts to the first subunit are shown in **a**. **c**, Ca^{2+}-independent interactions between the CaMBD and the CaM C-lobe. The van der Waals surface (cyan) is shown for the three distinct prongs that interact with this lobe, Ala 425, Leu 428 and Trp 432.

© 2001 Macmillan Magazines Ltd

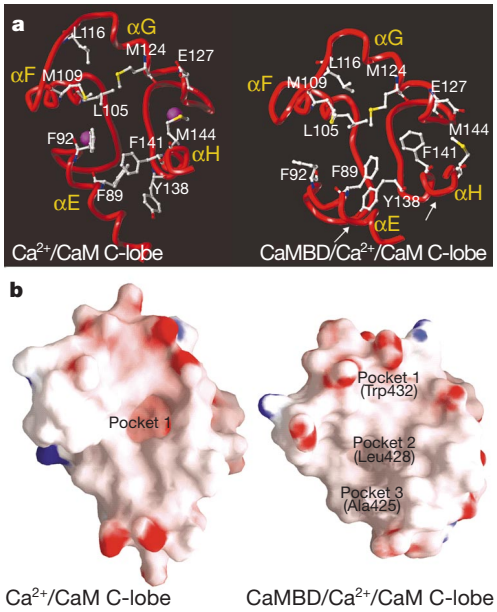


Figure 4 CaM C-lobe re-organization in the CaMBD/Ca^{2+}/CaM complex. **a**, Comparison of the C-lobe hydrophobic binding site between the Ca^{2+}/CaM structure[12] (left) and the CaMBD/Ca^{2+}/CaM structure (right). Note the striking core repacking (using αG as a fixed reference) in the CaMBD/Ca^{2+}/CaM structure leading to the creation of three smaller hydrophobic binding pockets. Also note the recruitment of Phe 89, buried in the Ca^{2+}/CaM complex, into the binding pocket of the CaMBD/Ca^{2+}/CaM complex. **b**, GRASP[30] surface representation of the structures shown in **a**, coloured for electrostatic potential (red, negative; blue, positive; white, neutral) showing surface of pockets.

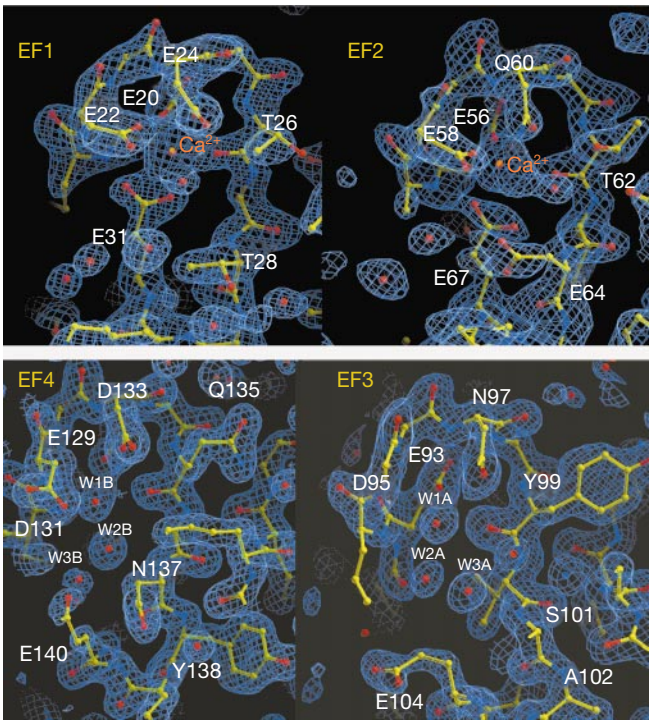


Figure 5 Omit map of N and C-lobe EF hand regions of CaMBD/Ca^{2+}/CaM complex. Composite $2F_o - F_c$ omit map (blue mesh) calculated in CNS[27] and contoured around the EF hands region of the N-lobe (top), contoured at 1.5σ, and C-lobe, contoured at 1.3σ (bottom). The final model is superimposed and represented as sticks with carbon, nitrogen, oxygen and Ca^{2+} shown in yellow, blue, red and orange, respectively. Figure generated using O[24].

results in an r.m.s. deviation of ~1.7 Å, confirming their strong similarity.

An important distinction between the NMR apo and CaMBD-bound structures of CaM is the slightly larger helical angle between αG and αH in the C-lobe of CaMBD-bound CaM (Fig. 3a), which is needed to create a more open binding pocket in the complex. The 'induced fitting' of these regions for complementation with the CaMBD is consistent with the finding that they are flexible in the NMR structures[13–15].

A striking result of the repacking of the C-lobe in the CaMBD complex is the separation of the primary hydrophobic binding pocket found in the Ca^{2+}-bound form (and used for binding other proteins) into several smaller and discrete pockets (Fig. 4a). Thus, instead of the typical one-pronged interaction, the CaMBD complements these pockets with three hydrophobic prongs formed by the side chains of Ala 425, Leu 428 and Trp 432 (Figs 2c and 4b). Consistent with fluorescence studies[4], Trp 432 is bound in a solvent-exposed pocket formed by the side chains of Glu 127, Met 124, Phe 141 and Met 144 (Fig. 2c). Unlike Trp 432, Leu 428 is completely buried and plugs a hydrophobic patch in the centre of the CaM C-lobe formed by Phe 89, Phe 92, Leu 105 and Met 109 (Fig. 2c).

Notably, Phe 89 is recruited from its buried position (in the Ca^{2+}-bound form of CaM[12]) into the CaMBD-binding pocket on the repacking of the C-lobe core. In this re-organization, Phe 89 switches its stacking partner from Phe 141 to engage the side chain of Tyr 138 (Fig. 4a; and Supplementary Information Fig. 2). The small Ala 425 side chain is critical in binding to the third pocket as it fits snugly into the small cleft formed by atoms of the C-lobe backbone and the side chains of Val 91 and Phe 92. These extensive contacts indicate a very stabile complex, consistent with functional analyses[4], in which the CaM C-lobe cannot be switched to the Ca^{2+}-bound conformation when bound to the CaMBD.

Consistent with the more apo-like conformation of the CaMBD-bound C-lobe, omit maps reveal that the C-lobe EF hands are Ca^{2+} free (Fig. 5). Instead of Ca^{2+}, clear density is observed for three water molecules that fill the expanded Ca^{2+}-binding sites in EF3 and EF4 (Fig. 5). Waters 1A and 1B are positioned close to the C-lobe Ca^{2+}-binding sites, perhaps indicative of Ca^{2+} ions at partial occupancy. This seems unlikely, however, given their lack of proper Ca^{2+} coordination, their average B factors (B_{ave} = 30 Å^2), which are close to the B_{ave} for solvents (Supplementary Information, Table 1), and the finding that refinement of these atoms as Ca^{2+} ions results in B factors greater than 100 Å^2.

Furthermore, there is a severe disruption of the C-lobe EF hand Ca^{2+}-coordination spheres, underscored by the large increase in Cα–Cα distances from Asn 97 to Glu 104 (EF3) and Asp 133 to Glu 140 (EF4), which are 11.3 Å and 11.4 Å, respectively, in the Ca^{2+}-bound conformation compared with 13.6 Å and 14.3 Å in the CaMBD complex. Moreover, the distances between the Ca^{2+}-coordinating side chain atoms of Asn 97 to Glu 104 and Asp 133 to Glu 140 are 4.7 Å in Ca^{2+}/CaM and 10–11 Å in the CaMBD/Ca^{2+}/CaM complex. Therefore, Glu 104 and Glu 140, which are essential for Ca^{2+} binding as they provide bifurcated contacts to theions, are effectively removed from their Ca^{2+}-binding positions (Figs 3b and 5). Indeed, these residues probably have pivotal roles in the Ca^{2+}-induced CaM conformational changes as, on Ca^{2+} binding, they must move into the Ca^{2+}-coordinating sphere—a conformational change that would lever their attached helices inward. The reverse process would account for the helical re-orientation that occurs on Ca^{2+} release.

The two fundamental features of K^+ channel function are K^+-selective permeation and gating. Structure–function studies[16–18] and the crystal structure of the KcsA potassium channel pore

© 2001 Macmillan Magazines Ltd

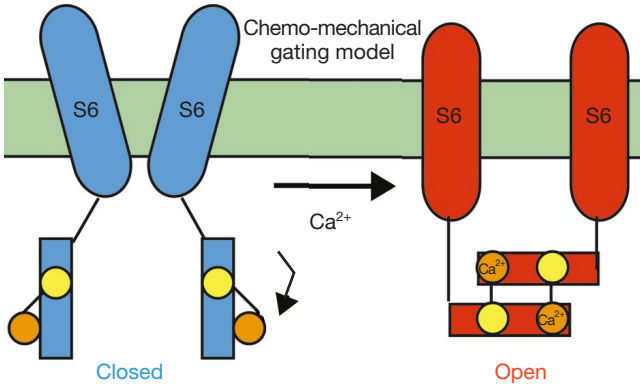


Figure 6 Cartoon of proposed chemo-mechanical gating model. Cutaway view showing two of the four subunits of the channel (the other two subunits would be in the foreground). For clarity only the S6 pore helices, the attached CaMBD and the linker (line) are shown. The CaM C-lobe and N-lobes are yellow and orange. Key to this model is biochemical data demonstrating that the CaMBD/CaM complex is monomeric in the absence and dimeric in the presence of Ca^{2+}. In this model, a resulting rotary movement (exaggerated in the figure) would result from the formation of the dimeric complex that would drive a rotation between the S6 helices to open the gate.

from *Streptomyces lividans*[19] have provided the basis for understanding K^+-selective permeation, a process that is probably conserved across all K^+ channels. The diverse array of K^+ channel subtypes are distinguished by what causes them to open and close, and this might reflect inherently different gating mechanisms.

Electron paramagnetic resonance spectroscopy studies on the proton-gated *S. lividans* K^+ channel suggest, however, that TM2 (corresponding to S6 in the 6-TM channels), which forms the channel pore, also functions as the gate. In this model, pH changes give rise to rotational/translational changes in the pore, most notably in the TM2 or S6 helix, that correspond to channel opening[20]. For voltage-dependent K^+channels, which also possess the 6-TM topology, mutational and metal crosslinking studies are consistent with a similar pore/gate mechanism[16–18]. Combined, these data suggest that although diverse, the metabolic signals that initiate gating in K^+ channels may work through a common mechanism involving S6 as the gate.

The intracellular CaMBD of SK channels resides immediately adjacent to S6 such that structural changes within or between CaMBD molecules might be directly transmitted to elicit gating. Dynamic light scattering (DLS) studies on the CaMBD/CaM complex (relative molecular mass (M_r 29K) of monomer complex 29,000) indicate that it is monomeric in the absence of Ca^{2+} (30K) and dimeric (61K) in the presence of Ca^{2+}. This has been confirmed by equilibrium sedimentation analyses (29K and 54K in the absence and presence of Ca^{2+}, respectively; Supplementary Information, Table 2). Combined with the structural data, these analyses suggest a possible gating mechanism in which Ca^{2+} binding to each CaM N-lobe would expose its hydrophobic patch and thus allow it to interact with an adjacent CaMBD monomer. As each N-lobe on adjacent monomers 'grabs' the other CaMBD C-terminal region, a rotary force would be created between them and transmitted to the attached S6 pore helices in the gate region (Fig. 6). In this chemo-mechanical model, two CaMBD dimers would serve as levers to drive opening of the channel (Fig. 6). □

Methods

Crystallization and data collection

We overexpressed and purified rat calmodulin and rSK2 CaMBD (see Supplementary Information) and ref. 21. The complex was dialysed into 10 mM Tris-HCl pH 7.5, 50 mM NaCl and 10 mM $CaCl_2$. For crystallization (hanging-drop vapour diffusion), 1 mM complex was mixed 1:1 with reservoir solution (1.25 M Li_2SO_4, 0.5 M $(NH_4)_2SO_4$, 0.1 M citrate pH 5.6) and equilibrated against 1 ml of reservoir solution at room temperature. The crystals are monoclinic, space group *C*2, and contain one CaMBD subunit, one CaM molecule and two Ca^{2+} ions in the asymmetric unit.

After structure determination, a suitable cryoprotectant (the crystallization solution supplemented with 25% glycerol) was used to dip the crystals for 1 min before plunging them into liquid nitrogen. X-ray diffraction data for the native and $PbCl_2$ derivative were collected at room temperature with an ADSC multiwire area detector and a RIGAKU RU200-H rotating anode X-ray generator with graphite monochromator, and processed with software provided by ADSC[22]. Data for the room temperature selenomethionine-CaMBD/Ca^{2+}/CaM and 1.98 Å cryo (~100 K) crystals were collected with an Raxis IV and a RIGAKU RU400-H rotating anode with Yale mirrors, and were processed with BIOTEX. We collected a high-resolution (1.60 Å) native data set at ~100 K at SSRL beamline BL 9-2 and processed it with MOSFLM.

Structure determination

The structure was solved by MIR with $PbCl_2$ and selenomethionineCaMBD/Ca^{2+}/CaM derivatives (Supplementary Information, Table 1). Two clear sites (shown to replace the two N-lobe Ca^{2+} ions) were observed in the Harker section. We refined heavy atom parameters, calculated MIR phases and carried out solvent levelling in PHASES[23]. Phases derived from the $PbCl_2$ derivative alone produced excellent maps, but because we could not account for about 20 residues of the CaMBD, we could not assign unambiguously the CaMBD residues. Therefore, selenomethionine-substituted CaMBD protein was overexpressed and purified as for the wild-type protein, except that 5 mM dithiothreitol was added to all buffers. The position of the selenium for Met 468, obtained from difference Fourier analysis, allowed unambiguous tracing[24] of the CaMBD with the exception of residues 395–412, for which there is no density. For CaM, clear side-chain density and the position of the linker were used to orient the highly symmetrical CaM molecule.

Model refinement

The initial structure, solved using room temperature data, was refined in TNT[25] to a limiting resolution of 2.35 Å resulting in an R_{factor} of 19.8% and an R_{free} of 28.0%. Initially, a cryo data set was collected in-house to 1.98 Å. The data from this crystal displayed significant differences in cell parameters (cryo cell constants: a = 77.6 Å, b = 66.3 Å, c = 64.6 Å and β = 93.9°) from the room temperature crystal. We therefore used the molecular replacement program EPMR[26] to position the room temperature model optimally. The EPMR solution had a correlation coefficient of 0.78 and the structure was refined in TNT.

A high-resolution (1.60 Å) cryo data set was collected subsequently at SSRL and the 1.98 Å cryo structure was used as the starting model in refinement, carried out in CNS, with this data. The model was subjected first to rigid body refinement followed by torsion-angle-simulated annealing at a starting temperature of 5,000 K (slow cooling) to 2.5 Å. After several rounds of model building in O[24] and refinement in CNS[27], during which progress was checked by monitoring the R_{free}, the refinement converged to an R_{cryst} of 22.8% and an R_{free} of 25.2% (Supplementary Information, Table 1). In addition to the protein complex, the current model contains 187 water molecules, 1 molecule of sulphate (located between CaMBD basic residues Arg 429, Lys 451, His 452 and Lys 455) and 5 histidines from the histidine tag. PROCHECK[28] revealed that 91.4% of residues are in most favourable regions, with 7.7% in additionally favoured regions. One residue, Gln 3 of CaM, is in the generously allowed region, and one residue, Lys 115 of CaM, is in the disallowed region. Although present, the density is weak for residues 1–9 of CaM, whereas the density for Lys 115 is excellent and consistent with the modelled ϕ/y angles.

DLS and sedimentation equilibrium

The oligomerization state of the purified SK2 CaMBD/Ca^{2+}/CaM complex with (2–10 mM) and without (5 mM EGTA) Ca^{2+} in 10 mM Tris, pH 7.5, 50 mM NaCl was examined by DLS and sedimentation equilibrium at concentrations of ~3 mg ml^{-1} (see Supplementary Information). These measurements gave molecular weights for the Ca^{2+}-free and -bound forms of the complex of 30,000 ± 6,000 and 61,000 ± 3,000 by DLS and 29,000 ± 5,000 and 54,000 ± 5,000 by sedimentation equilibrium, respectively (Supplementary Information, Table 2).

Received 18 October 2000; accepted 8 February 2001.

1. Köhler, M. *et al.* Small-conductance, calcium-activated potassium channels from mammalian brain. *Science* **273,** 1709–1714 (1998).
2. Vergara, C., Lattore, R., Marrion, N. V. & Adelman, J. P. Calcium-activated potassium channels. *Curr. Opin. Neurobiol.* **8,** 321–329 (1998).
3. Xia, X.-M. *et al.* Mechanism of calcium gating in small-conductance calcium-activated potassium channels. *Nature* **395,** 503–507 (1998).
4. Keen, J. E. *et al.* Domains responsible for constitutive and Ca^{2+}-dependent interactions between calmodulin and small conductance Ca^{2+}-activated potassium channels. *J. Neurosci.* **19,** 8830–8838 (2000).
5. Ikura, M. *et al.* Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science* **256,** 632–638 (1992).
6. Ikura, M., Barbato, G., Klee, C. B. & Bax, A. Solution structure of calmodulin and its complex with a myosin light chain kinase fragment. *Cell Calcium* **13,** 391–400 (1992).
7. Meador, W. E., Means, A. R. & Quiocho, F. A. Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin-peptide complex. *Science* **257,** 1251–1255 (1992).

© 2001 Macmillan Magazines Ltd

8. Osava, M. *et al.* A novel target recognition revealed by calmodulin in complex with Ca^{2+} calmodulin dependent kinase kinase. *Nature Struct. Biol.* **6,** 819–826 (1999).
9. Elhorst, B. *et al.* NMR solution structure of a complex of calmodulin with a binding peptide of the Ca^{2+} pump. *Biochemistry* **38,** 12320–12326 (1999).
10. Rhoads, A. R. & Friedberg, F. Sequence motifs for calmodulin binding. *FASEB J.* **11,** 331–339 (1997).
11. Babu, Y. S. *et al.* Three-dimensional structure of calmodulin. *Nature* **315,** 37–40 (1985).
12. Chattopadhyaya, R., Meador, W. E., Means, A. R. & Quiocho, F. A. Calmodulin structure refined at 1. 7 Å resolution. *J. Mol. Biol.* **228,** 1177–1192 (1992).
13. Zhang, M., Tanaka, T. & Ikura, M. Calcium-induced conformational transition revealed by the solution structure of the apo calmodulin. *Nature Struct. Biol.* **2,** 758–767 (1995).
14. Kuboniwa, H. *et al.* Solution structure of calcium-free calmodulin. *Nature Struct. Biol.* **2,** 768–776 (1995).
15. Finn, B. E. *et al.* Calcium-induced structural changes and domain autonomy in calmodulin. *Nature Struct. Biol.* **2,** 777–783 (1995).
16. Liu, Y., Holgrem, M., Jurman, M. E. & Yellen, G. Gated access to the pore of a voltage-dependent K^+ channel. *Neuron* **19,** 175–184 (1997).
17. del Camino, D., Holgrem, M., Liu, Y. & Yellen, G. Blocker protection in the pore of a voltage-gated K^+ channel and its structural implications. *Nature* **403,** 321–325 (2000).
18. Holmgren, M., Shin, K. S. & Yellen, G. The activation gate of a voltage-gated K^+ channel can be trapped in the open state by an intersubunit metal bridge. *Neuron* **21,** 617–621 (1998).
19. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280,** 69–77 (1998).
20. Perozo, E., Cortes, D. M. & Cuello, L. G. Structural rearrangements underlying K^+-channel activation gating. *Science* **285,** 73–78 (1999).
21. Fragioni, J. V. & Neel, B. G. Solubilization and purification of enzymatically active glutatione S-transferase (pGEX) fusion proteins. *Anal. Biochem.* **210,** 179–187 (1993).
22. Howard, A. J., Nielson, C. & Xuong, N. H. Software for a diffractometer with multiwire area detector. *Methods Enzymol.* **114,** 452–472 (1985).
23. Furey, W. B. & Swaminathan, S. PHASE-95: a program package for the processing and analysis of diffraction data from macromolecules. *Methods Enzymol.* **277,** 590–620 (1997).
24. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
25. Tronrud, D. E., TenEyck, L. F. & Matthews, B. W. An efficient general purpose least-squares refinement program for macromolecular structures. *Acta Crystallogr. A* **43,** 489–501 (1985).
26. Kissinger, C. R., Gehlaer, D. K. & Fogel, D. B. Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr. D* **55,** 484–491 (1999).
27. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
28. Laskowski, R. A., MacArthur, M. W. & Thorton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).
29. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24,** 946–950 (1991).
30. Nicholls, A., Sharp, K. & Honig, B. H. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11,** 281–296 (1991).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

M.A.S. is a Burroughs Wellcome Career Development Awardee. This work was supported by grants from the NIH, the Muscular Dystrophy Association, and the Human Frontiers of Science Foundation. Intensity data collected at the Stanford Synchrotron Radiation Laboratory (SSRL) was carried out under the auspices of the SSRL biotechnology program, which is supported by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program, and by the Department of Energy, Office of Biological and Environmental Research.

Correspondence and requests for materials should be addressed to M.A.S. (e-mail: schumacm@ohsu.edu).

Coordinates have been deposited with the Protein Data Bank under accession code 1G4Y.

erratum

Increase in greenhouse forcing inferred from the outgoing longwave radiation spectra of the Earth in 1970 and 1997

John E. Harries, Helen E. Brindley, Pretty J. Sagoo & Richard J. Bantges

Nature **410,** 355–357 (2001).

In Fig. 1a of this paper, the labels for the two curves were inadvertently switched. The grey curve represents IMG and the black curve represents IRIS. □

© 2001 Macmillan Magazines Ltd

Be prepared

The theme this week is 'DNA/RNA preparation' and all that that entails.

Bactozol

Molecular Research Center, Inc.
www.mrcgene.com

A novel solution

The Bactozol kit contains Bactozyme, a concentrated (×10) solution for lysis of bacterial cell walls, and DNAzol for isolation of DNA from the bacterial lysate. It is suitable for isolation of DNA from a broad spectrum of Gram-negative and Gram-positive bacteria. Bactozyme remains stable at room temperature, precluding the need for daily enzyme preparation. The kit provides sufficient reagents for 125 isolation procedures, each yielding up to 40 mg of DNA.

Primer D'Signer

IBA www.iba-go.de

Freeware for Strep*-tag users*

This Windows-based software program automatically develops primers for PCR cloning of open reading frames into pASK-IBA (*Strep*-tag) expression vectors. It checks the reading frame, looks for gene internal restriction sites and provides an optimal cloning strategy. The software can be downloaded free from the company's website.

GELase

Epicentre www.epicentre.com

Gentle, high-yield preparation

GELase solution recovers intact DNA and RNA from low melting point (LMP) agarose gels following electrophoresis in TAE, TBE, NOPS or phosphate buffers in less than 10 minutes and using only three units of enzyme. The procedure is designed to be gentle, and recovery yields approach 100%. Purified nucleic acid can be rapidly recovered from the digested gel solution by ammonium acetate/ethanol precipitation. Recovered DNA is then ready to be used in restriction mapping, cloning, labelling, DNA sequencing, transformation of YAC, BAC, cosmid and plasmid vectors, micro-injections and amplification. GELase is available at 1 U per ml for standard reactions and 0.2 U per ml for multiple small gel samples or extended digestion times.

SPE Dry Dual

Jones Chromatography www.joneschrom.com

High and dry

This dual 96- and 384-well plate evaporator features side-by-side needle assemblies that can dry samples in two microplates simultaneously. The use of heated gas, typically nitrogen, delivered through needle assemblies located above and below the microplate, gives claimed concentration speeds five times those achieved with other methods. The assemblies consist of 96 or 384 individual stainless steel needles that evenly distribute the heated gas into the microplates, enhancing evaporation of aqueous solvents. The plate holder can be adjusted to any height from the upper needle assembly to accommodate a variety of plate sizes and sample volumes. The evaporator is suitable for high-throughput applications in SPE, combinatorial chemistry and DNA synthesis.

SPE Dry Dual: suited to DNA synthesis.

Nucleon PhytoPure

Tepnel www.tepnel.co.uk

Get a grip on slippery samples

The Nucleon PhytoPure kit was developed to overcome problems caused by 'slimy' samples and poor enzymatic analysis of DNA from plant samples caused by the presence of polysaccharides in the extracted DNA. The protocol, which takes less than one hour, uses a resin that binds with polysaccharides, thus removing them from the sample and yielding pure DNA. It can be used with a wide range of fresh, frozen or freeze-dried plant materials and can also be used with other carbohydrate-containing samples such as nematodes and snails to remove the problems associated with mucopolysaccharides. The resulting DNA is suitable for downstream processes including RFLP, RAPD and AFLP analyses.

Mismatch substrate

Trevigen www.trevigen.com

DNA repair is a Mug's game

Mug (mismatched uracil glycosylase) is a DNA damage and repair enzyme first identified in *E. coli* based on homology to human thymine-DNA glycosylase. Although Mug protein can remove a uracil base from within a U·G mismatch, the role of Mug *in vivo* is believed to involve the repair of 3,N^4-ethenocytosine-G mismatches. Trevigen provides purified Mug in 100- and 500-unit sizes with an appropriate reaction buffer for *in vitro* testing.

Reverse-iT

ABgene www.abgene.com

Hot-start kit

ABgene has devised a new system for one-step, hot-start RT-PCR by combining the Reverse-iT blend with Thermo-Start DNA polymerase in one kit. The Reverse-iT blend maximizes the sensitivity and robustness of the RT step, while the Thermo-Start removes the possibility of nonspecific primer extension prior to PCR, eliminating spurious bands. The kit is supplied as a ×2 master mix containing the ThermoStart polymerase, optimized buffer, magnesium chloride and dNTPs. A one-step

ADVERTISEMENTS

DIVERSE SMALL MOLECULES READY FOR SCREENING

CHEM BRIDGE™

10K–50K Compounds in Plates

• Next Day Delivery*
• Competitively Priced
*(Some Restrictions Apply)

Chembridge Corporation • 800.964.6143
858.451.7400 • 858.451.7401 FAX • E-mail: chem@chembridge.com

www.chembridge.com

© 2001 Macmillan Magazines Ltd

master mix format reduces the number of pipetting steps. The Reverse-iT blend is supplied in a separate tube so that no-RT controls can be performed.

UV-transparent microplates

BD Biosciences www.bd.com

You can see through these

Available in 96- and 384-well formats, Falcon UV-transparent microplates allow quantification of DNA or RNA without the use of quartz. With low UV absorbance down to 240 nm, typical measurements for DNA concentration at 260–280nm can be easily monitored. Both 96- and 384-well formats feature SBS standard footprints.

ExpressDirect

Pierce www.piercenet.com

mRNA capture and RT system

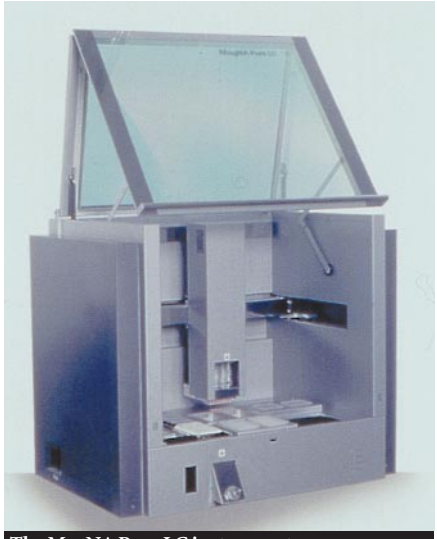
A complete system for cell lysis, mRNA capture and reverse transcription for RT-PCR, ExpressDirect uses a two-step lysis procedure that selectively disrupts only the outer cell membrane so that processed mRNA — with no contaminating genomic DNA — is released from the starting cell or tissue sample.The mRNA is then captured directly from the cell lysate by hybridization of oligo(dT) that is covalently linked onto the wells of a 0.2 ml, 96-well PCR plate. First-strand cDNA synthesis of the captured mRNA is performed in the same tube, reducing the risk of contamination. A flexible microtube plate format allows the user to perform 1–96 reactions.

MagNA Pure LC II

Roche www.biochem.roche.com/magnapure

DNA researchers will find these attractive

Two new reagent kits — the MagNA Pure mRNA Isolation Kit II and the DNA Isolation Kit II — are now available for use with Roche's MagNAPure LightCycler instrument. The MagNA Pure system is based on magnetic bead technology and can isolate nucleic acids from a variety of human and animal tissues, including liver, kidney, spleen and mouse tail. Frozen tissue samples (1–10 mg) are homogenized and directly entered into the automated isolation process. The system can process 32 samples in 90 minutes. The kits contain sufficient reagent for 192 isolations.



The MagNA Pure LC instrument.

LCM/DNA extraction kit

Arcturus www.arctur.com

Optimized for LCM-captured cells

Laser capture microdissection (LCM) technology allows researchers to localize, extract and analyse specific single cells or groups of cells microdissected from solid tissue samples, cytological smears and cell cultures. This kit is based on the proteinase K extraction method and allows complete extractions from a variety of tissue types. The extracted DNA can be taken directly from digestion to gel electrophoresis, PCR, quantitative real-time PCR and other molecular analyses with no additional clean-up steps.

S5700

Beckman Coulter www.beckmancoulter.com

A complete package for DNA research

This rotor, for use with the Allegra 25R bench centrifuge, holds five standard and two deep-well plates per bucket, and runs widely-used commercial kits for DNA sample preparations, including those manufactured by Promega and QIAGEN. The Allegra 25R system handles samples from pelleting of starting material through to DNA isolation and post-reaction clean-up.

Fluorlink CyDye

Amersham Pharmacia www.apbiotech.com

A sizeable saving

FluoroLink CyDyes, a range of fluorescent dyes based on cyanine fluors and used for labelling biological molecules, are now available in 'value pack' sizes. The dyes are suitable for multiple colour detection in microarray and other fluorescence-based genome analysis techniques. The value packs offer a choice of dCTP or dUTP, each containing 5 × 25 nmol of both Cy3- and Cy5-labelled nucleotides.

cDNA Archive Kit

Applied Biosystem appliedbiosystems.com

Store RNA as cDNA

This new High-Capacity cDNA Archive Kit has been launched to satisfy the demand for the large quantities of cDNA which gene expression studies require. The kit converts up to 10 μg of RNA to single-stranded cDNA in a single 100-μl reaction, larger quantities of cDNA than are obtainable with traditional RT kits or enzyme/dNTP/buffer combinations. Suggested applications include gene expression studies, long-term storage, and other reverse transcriptase (RT) applications.

These notes are compiled in the Nature *office from information provided by the manufacturers.*

ADVERTISEMENTS

BACs • YACs • PACs
cDNAs

We provide clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

CUSTOMISED
monoclonal antibodies with
GUARANTEE

BioGenes
Gesellschaft für Biopolymere mbH

Internet www.biogenes.de
Berlin ¥Germany
Phone +49 (0)30-65 76 23 96

Polyclonal Antibodies
& Custom Peptides

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

© 2001 Macmillan Magazines Ltd

Contacts
Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London
The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:
Leonie Welss (4954)
European Manager:
Nevin Bayoumi (4978)
European Sales Consultant:
Jonathan Bull (4973)

UK/ RoW/ Ireland:
Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)
Holland/ Italy: Nevin Bayoumi (4978)
Scandinavia: Sille Opstrup (4994)

Production Manager: Billie Franklin
To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

International
Advertising Coordinator:
Laura Pearson (4977)
Naturejobs web development:
Tom Hancock
Naturejobs online production:
Ben Lund

European Satellite Offices
France/ Belgium:
Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Spain/ Portugal:
Amanda Wren
Tel + 34 91 3913835
Fax + 34 91 319 0091
e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:
Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York
345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: classified@natureny.com

US Sales Director: Ben Crowe
US Sales Manager: Peyton Mason

US Advertising Coordinator:
Corrisa Salzman

Japan Head Office, Tokyo
Shin-Mitsuke Building (4F),
3-6 Ichigaya Tamachi Sinjuku-Ku,
Tokyo 162
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejpn.com
Japan Manager: Kate Cowan

naturejobs
nature.com

Physics comes full circle

Just eight years ago, reports were bemoaning the plight of the young physicist. There were no jobs in academia, industry or government. These reports seemed to discourage students from pursuing a PhD in physics. Students who started a PhD anyway — despite the best available advice — are now emerging to a markedly changed job market in which all sectors are seeking physicists.

In this month's *Physics Today*, authors of a 1993 American Institutes of Physics article reflect on their earlier predictions, and draw some new conclusions from the present data (http://physicstoday.org/pt/vol-54/iss-4/p36.html#bio). US and European laboratories are now having a difficult time getting nationals to take up physics. It is also increasingly difficult to recruit physicists from Russia and China — traditionally good sources of talent.

The change in fortune serves as a cautionary tale about making long-range predictions of employment in any field, no matter how good the available data. The latest report could be in as much danger of being out of date as the one it replaces. After the report was published, the latest US budget proposal by President George W. Bush proved to be less than physics-friendly. And many of the technology companies that grabbed physicists when academic and government opportunities were scarce seem to be faltering in the stock market.

Of course, even if the stock market crashes and Bush gets his way with the budget, there will still be jobs for physicists. Many academic physicists are approaching retirement, and other disciplines, such as structural biology, are increasingly dependent on physics. So, even considering the worst-case scenarios, it may be prudent to find creative ways to attract the best talent into physics programmes, rather than scare them away.

Paul Smaglik
Naturejobs editor

Contents

MOVERS

Cancer research scientist moves south, plant biologist becomes dean, Chinese ministry gets new boss, and more Back page

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific job market

FOCUS
SPOTLIGHT
RECRUITMENT
SCIENTIFIC ANNOUNCEMENTS
SCIENTIFIC EVENTS

© 2001 Macmillan Magazines Ltd

BIOLOGY

David Thurston, director of a centre at the University of Nottingham funded by the Cancer Research Campaign (CRC), is relocating to the University of London in July, following a tobacco controversy. The CRC had agreed to raise £1.5 million (US$2.1 million) for the group's new research facility. But the organization balked after the university accepted £3.8 million from British American Tobacco, to build a new department within the business school called the Institute for Corporate and Social Responsibility. Although the CRC did not threaten Thurston's existing funding, he was concerned that future support for other members of his team could be jeopardized. The University of London offered Thurston renovated lab space in the School of Pharmacy. He is also relocating his spin-off company, SpiroGen, to London.

Sharon Long, a researcher who studied nitrogen fixation in plants, has been appointed dean of Stanford University's School of Humanities and Sciences. Long is the William C. Steere, Jr/Pfizer Professor in Biological Sciences; professor, by courtesy, in the Department of Biochemistry; and investigator at the Howard Hughes Medical Institute (HHMI). She succeeds Malcolm Beasley, who will step down in August after three years in post. The school is the largest of Stanford's seven and encompasses the core humanities, fine arts, languages and literatures, social sciences, mathematics and the natural sciences. One of Long's goals is to foster more collaborations between the mathematical and experimental sciences. When she assumes her new role as dean, Long will step down from her position on the Bio-X executive committee and as an HHMI investigator. She will continue her research on bacterial–plant symbiosis but with fewer staff.

The Stowers Institute for Medical Research, a US$200 million facility which formally opened this month in Kansas City, Missouri, recently appointed four new scientists, all of whom will join the institute this summer. The four new appointments, along with two announced in December, effectively double the number of scientific teams working at the cancer research institute. **Arcady Mushegian** will head the institute's bioinformatics programme. Mushegian currently leads the bioinformatics programme at Akkadix Corporation, an agricultural biotechnology company based in La Jolla, California. **Kent Golic** will lead a team in understanding gene regulation in *Drosophilia*. He is currently professor of genetics at the University of Utah, Salt Lake City. **Paul Trainor** will study interactions between different kinds of tissues in mice. Trainor is currently at the National Institute for Medical Research in Mill Hill, Britain. **Chunying Du** will investigate apoptosis in cancer. Du is now a Howard Hughes Medical Institute postdoc at the University of Texas Southwestern Medical Center in Dallas. Bill Neaves, president and chief executive of the Stowers Institute, aims to have 50 to 60 scientific teams, each of about 10 people, including technicians, in place by 2010.

Chunying Du

POLICY

Xu Guanhua, a remote-sensing expert who has been in charge of promoting China's high-tech enterprises, will head the country's Ministry of Science and Technology. He succeeds **Zhu Lilan**, who assumes a top legislative post within the National People's Congress. Xu will direct a growing science and technology budget that reached US$6.5 billion in 1999. He oversees state-run scientific institutes, including the Chinese Academy of Sciences (CAS), as well as funding for key basic research projects, high-technology development, scientific infrastructure and international collaborations. A native of Shanghai, Xu was trained as a forestry scientist and spent 30 years working for the Chinese Academy of Agricultural and Forestry Sciences before moving to the CAS. In 1992, he was named an academician, a high honour in Chinese science. Xu is credited with helping to develop the country's remote-sensing industry with global information system instruments.

Arcady Mushegian

PHYSICS

Laser physicist **John Hegarty** will begin a 10-year term as provost of Trinity College in Dublin this August. He has promised to encourage innovation in teaching, increase the number of students from under-represented groups and to promote excellence in research. He has both academic and industrial experience; after spells at the University of Wisconsin, Madison, and Bell Laboratories in the US, he moved to Trinity as professor of laser physics in 1986. Since then he has also served as head of the physics department and dean of research. He also established Optronics Ireland in 1989.

Sharon Long

CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: naturejobseditor@naturedc.com.

© 2001 Macmillan Magazines Ltd